Information Sciences 741 (2026) 123274

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

VEUCTOR: Training and selecting best vector space models from online job ads for European countries☆

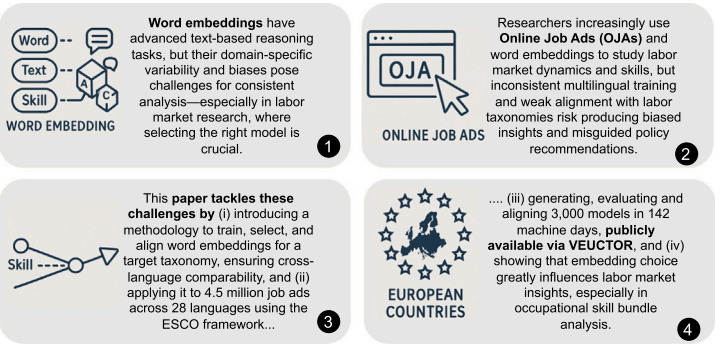
Emilio Colombo[a], Simone D'Amico[b], Fabio Mercorio[b,*], Mario Mezzanzanica[b]

[a] Dept of Economics, Università Cattolica del Sacro Cuore, Milan, Italy
[b] Dept of Statistics and Quantitative Methods - University of Milan-Bicocca, Milan, Italy

HIGHLIGHTS

- We present, formalise, and implement a multilingual methodology to train, select, and align word embedding models using the ESCO taxonomy across 28 European countries.
- We generate and evaluate over 3000 embedding models trained on 4.5 million online job advertisements in the frame of an EU Project, using a benchmark-driven approach to optimize semantic alignment.
- We release VEUCTOR, a tool that provides access to the best-performing and aligned embeddings, enabling reuse and supporting third-party labor market analyses.
- We show that the choice of embedding significantly affects occupational skill bundles and, consequently, labor market analysis outcomes.
- We enable reproducible and cross-country labor market intelligence by standardizing model development and alignment across diverse languages and corpora.

GRAPHICAL ABSTRACT



☆ This paper is partially supported within the research activity of a grant entitled *"PILLARS - Pathways to Inclusive Labour Markets"* - under the call H-2020 TRANSFORMATIONS 18-2020 "Technological transformations, skills, and globalization - future challenges for shared prosperity", grant agreement NUMBER 101004703 - PILLARS https://www.h2020-pillars.eu/

* Corresponding author.
Email addresses: emilio.colombo@unicatt.it (E. Colombo), simone.damico@unimib.it (S. D'Amico), fabio.mercorio@unimib.it (F. Mercorio), mario.mezzanzanica@unimib.it (M. Mezzanzanica).
URL: mercorio.com (F. Mercorio).

https://doi.org/10.1016/j.ins.2026.123274
Received 12 April 2025; Received in revised form 19 February 2026; Accepted 19 February 2026
Available online 21 February 2026

ARTICLE INFO

Keywords:
Word embedding
Machine learning
Labor market
NLP

ABSTRACT

Over the last decade, word embeddings have enabled machines to represent words and sentences as vectors, enabling researchers to reason on text for tasks like semantic similarity, contextual understanding, machine translation, etc. However, the synthesis of embeddings involves domain-specific parameters that affect semantic accuracy and contextual relevance, often leading to unpredictable biases and inconsistent comparisons. This issue is particularly relevant in labor market analysis, where different embeddings yield varying results, making the selection of the most appropriate model a key element.

This paper addresses these challenges by (i) proposing a methodology to train, select, and align vector space models for a target taxonomy, ensuring comparability across dimensions and languages; (ii) applying this approach to 4.5 million job ads in 28 languages, aligning country-specific embeddings using the ESCO taxonomy; (iii) generating over 3000 models over 142 machine days, making the best-performing ones publicly available via `VEUCTOR`; and (iv) showing how model choice significantly impacts labor market analysis, revealing substantial variations in occupational skill bundles across embeddings.

1. Introduction and motivation

As labor markets rapidly evolve, harnessing the power of data to decode workforce skills has never been more crucial. The European Skills Agenda[1] highlights the strategic importance of skills in driving economic competitiveness, yet measuring and identifying them remains a daunting task. In 2025, the European Union published the "Union of Skills",[2] to clarify that investing in people's skills is key for Europe's economic competitiveness, resilience, and social cohesion, thus suggesting the development of a comprehensive strategy addressing skills shortages, gaps, and mismatches, ensuring that individuals and businesses have the necessary skills for success in the evolving global economy. Notably, the Commission plans to consolidate all labor-related data - including labor shortages and surpluses reports and skills online job ads analysis tools - from agencies such as Eurostat, Cedefop, Eurofound, and the European Labor Authority into a unified data lakehouse. This, in turn, would enable real-time intelligence and skills forecasting for the Union to support policy and decision-making.

Though these initiatives highlight the importance of skills intelligence in analyzing labor market dynamics, the concept of "skill" remains ambiguous and often subject to misinterpretation, as extracting, recognizing, and analyzing them presents two key challenges: The first is the development of a valid taxonomy for skill classification. The second is the identification and extraction of skills from available sources. O*NET has addressed the first challenge in the US and ESCO in Europe, which now provide a consistent and robust taxonomy available to researchers and analysts. The second challenge, however, is far more complex. Much of the valuable, skill-related information is buried in unstructured text - such as online job advertisements - requiring advanced language models to extract meaningful insights. At the core of this process lies a crucial element: word embeddings, which unlock hidden patterns in language and transform raw text into actionable intelligence, as they enable machines to reason with text.

Generally speaking, Word embedding can be seen as a technique in natural language processing that represents words as dense vectors in a continuous vector space, capturing their semantic relationships and contextual meanings to improve machine understanding of language. As stated above, embeddings are crucial for identifying skills within a text. However, the process of creating embeddings is often arbitrary, leading to potentially biased outcomes. Different embedding models can produce significantly different results, making the selection of the right model essential; in fact, there is no established standard for selecting embeddings. This paper tackles this issue by showing that different embedding choices can lead to dramatically different results and by proposing a novel framework for identifying the best embedding model for skill extraction. More specifically, we provide three main contributions to the literature.

Contribution. First, we define and implement a methodology to train and select vector space models that best fit a specific target taxonomy across multiple dimensions. We then align them to enable direct comparisons between models. The alignment process is particularly important in multi-language contexts such as the EU, where language-specific issues may affect embeddings. Although the methodology is domain-independent, it is applied here within the field of labor and skill intelligence. It has been trained over 4.5 million job advertisements in 28 languages to build vector space models, using ESCO[3] as the target taxonomy and aligns country-specific embeddings to allow for cross-country comparisons.

Second, we synthesize over 3000 embedding models for all 27 EU countries, plus the UK, identifying the best- and worst-performing embeddings with a total training time of 142 machine days. These models are aligned to the EU benchmark (i.e., the UK) and are made available to the community as an off-the-shelf Python tool, namely `VEUCTOR`, allowing easy adoption and improving the reproducibility and comparability of different labor market analyses for the scientific community.

[1] https://ec.europa.eu/social/main.jsp?catId=1223&langId=en.

[2] COM(2025) 90 final. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_657.

[3] The ESCO taxonomy provides a structured representation of Skills, Competencies, Qualifications, and Occupations relevant to the European labor market. The European Commission has devised it to be a dictionary for the labor market in 27 + 1 countries and 32 languages. https://esco.ec.europa.eu/.

Third, we demonstrate the relevance of this work for real-world labor market analysis. We construct the occupational skill bundles derived from different embedding models, and we show that there is a large difference in the skill bundles, highlighting how the choice of embedding model can significantly impact the conclusions drawn from skill analysis. This finding is even more significant following the number of highly relevant publications that use OJAs to conduct skill analysis.

More generally, although our work is applied to the EU labor market, our results can be generalized and extended to several other domains. The availability of a large amount of unstructured information contained in texts has led to a proliferation of tools, analyses, and research that process and analyze such data. For example, in Economics, there is a vast literature that has analyzed Central banks' communications and their effects on markets and expectations (see, for example, [6,24]). What we show in this paper is that the method of information extraction is not neutral, and the results can be highly dependent on it. This problem cannot be solved simply by more transparency, but must be accompanied by an optimization analysis such as the one we propose.

Roadmap. The methodology presented in this paper follows a structured sequence of steps, which we summarize as follows:

Establishing a Benchmark: To evaluate the optimality of different word embeddings, a reference framework is needed for comparison. Since our embeddings are developed from a sample of Online Job Advertisements, we use the ESCO taxonomy as the natural benchmark.

Assessing Proximity to the benchmark: Once the benchmark is defined, a suitable methodology is required to measure the alignment of different embeddings with ESCO. We employ the Hierarchical Semantic Similarity (HSS) method for this analysis.

Addressing Multilingual Variability: Given the multilingual nature of our dataset, language differences can introduce distortions in embedding comparisons. To mitigate this, we implement an embedding alignment technique that ensures cross-language consistency.

Quantifying Differences: After evaluating the optimality of different embeddings, we measure the extent of variation between them. Specifically, we compare the best and worst embeddings by constructing a similarity metric for skill bundles. Our results indicate significant differences between data generated by the most and least optimal embeddings.

Providing data: Given the large differences in outcomes following different embeddings, we make available to the research community a data tool, `VEUCTOR`, which contains the codes and data for the best- and worst-generated embeddings, along with their optimality scores.

The next section provides a more detailed description of the above-mentioned steps.

2. Preliminaries

2.1. Online job advertisements for labor and skill intelligence

In recent years, the role of specialized web portals and services in intermediation has grown exponentially. This surge has given rise to the concept of Labor Market Intelligence (LMI), which involves leveraging AI algorithms and frameworks to analyze labor market data and support data-driven decision-making (see, e.g., [21,38,44]). In this context, the ability to monitor, analyze, and understand labor market changes both (i) in a timely manner and (ii) at a highly granular geographical level has become increasingly significant. Recent studies have further pushed this boundary by using LLMs to assess how emerging technologies affect the labor market, such as evaluating the exposure of occupations and skills to AI [13].

Online Job Advertisements (OJAs) possess these features and have become increasingly important for academic research and the development of innovative statistics in recent years. Academic research has exploited the granularity of OJAs to analyze labor market concentration [2,3,11,25], but most importantly has leveraged the information contained in the text of the advertisement to analyze firms' skill requirements [9,12,22,23,33]. As stated in the introduction, information extraction from textual data relies crucially on Word embeddings; moreover, most research is conducted in the US due to the lack of available methodological approaches that can deliver robust results in a multilingual setting. Our paper addresses both issues by providing a robust methodology to identify the best Word embedding in a multi-language context.

This is relevant not only for research purposes but also for the production of statistical data supporting skill-related policies. In the European context, in 2016, the European Commission issued the communication "A New Skills Agenda for Europe"[4] highlighting a number of actions and initiatives aimed at equipping the European labor force with the skills of the future. In support of this initiative, the EU agency Cedefop subsequently teamed up with Eurostat to develop a system capable of collecting and classifying online job vacancies for the entire EU, covering all 28 Member States and the Union's 32 languages.[5]

The result of this effort is the Web Intelligence Hub (WIH), which has collected OJAs from more than 1000 sources in Europe since 2019. Several results from this effort have been published in studies such as [8,12]. This study leverages this initiative by utilizing the knowledge base compiled by the WIH to train and optimize word embeddings.

[4] COM(2016) 381/2, available at https://migrant-integration.ec.europa.eu/sites/default/files/2020-07/SkillsAgenda.pdf.

[5] https://www.cedefop.europa.eu/it/about-cedefop/public-procurement/real-time-labour-market-information-skill-requirements-setting-eu.

2.2. Word embeddings

The task of learning meaningful representations for words and documents from a text corpus is fundamental in Natural Language Processing (NLP). Over the years, various approaches have been developed to transform words into numerical vectors, enabling their use in Machine Learning (ML) models. These representations encode words in a way that captures their semantics and contextual relationships within a text corpus, allowing words to be manipulated mathematically and facilitating operations such as similarity measurement, clustering, and classification. Early techniques for word representation, known as frequency-based models, relied on counting the occurrences of words in documents. While effective in some applications, these models present notable limitations. First, they produce high-dimensional and sparse representations, making them computationally expensive and memory-intensive. More importantly, they fail to capture semantic relationships between words, as they treat each term as an independent entity, ignoring contextual similarities and linguistic structures.

To overcome these limitations, distributional approaches were introduced based on the principle that words appearing in similar contexts tend to have related meanings. This led to the development of dense vector representations, known as word embeddings, which encode words in continuous, low-dimensional spaces, capturing both semantic and syntactic relationships.

A key step in this area came with the introduction of neural embedding models such as Word2Vec [35]. This method learns word representations using a shallow neural network and can be trained through two alternative architectures: *Continuous Bag of Words* (CBOW) and *Skip-Gram* (SG). CBOW predicts a target word based on its surrounding context, resulting in fast training and good performance on large corpora. Skip-Gram instead predicts context words given a target word, enabling it to better capture relationships between rare words by generating more training examples from a given corpus. These innovations significantly improved the modelling of semantic similarity and paved the way for more advanced word-embedding techniques.

Bojanowski et al. [7] developed *fastText*, an extension of Word2Vec that incorporates sub-word information. Unlike Word2Vec, which represents each word as a single vector, fastText models a word as the sum of its character *n*-gram vectors. This enriched representation offers several advantages. By leveraging sub-word information, fastText can better handle rare words, typos, and morphologically rich languages.

Motivating Example of Word Embeddings for Labor Market

A graphical representation of a word embedding model is shown in Fig. 1, trained on millions of online job advertisement titles. The map highlights existing concepts from the ESCO taxonomy (represented by empty shapes) and alternative labels (terms that emerge from online data and are strongly related to ESCO concepts but are not yet included in ESCO).

The embedding model effectively *"encodes"* words with similar meanings within the context of the labor market. For example, while a *data engineer* and a *data scientist* are both categorized as sub-concepts under the *2511: System Analyst* ISCO group in ESCO, their real-world roles differ significantly - something any computer scientist would recognize. In practice, a data engineer often aligns more closely with *2521: Database Designers and Administrators* than with its theoretical ISCO group. Conversely, there are instances where the taxonomy accurately reflects real labor market demands. A clear example is *3521: Broadcasting and Audio-Visual Technicians*, which forms a tight cluster on the map, indicating a strong alignment between de-facto and de-jure labor market occupations. Similarly, *3513: Computer Network and Systems Technicians* also demonstrates this consistency, albeit to a slightly lesser extent. Notably, representing words as semantic vectors in vector space enables the use of algebra to perform operations over concepts. This means one can perform the following vector operation:

$$\vec{v}_{it\_security\_manager} - \vec{v}_{security} + \vec{v}_{data} = \vec{v}_{data\_quality\_manager}$$

This highlights the capability to reason mathematically over words and their semantic relationships using vector algebra. Consequently, it underscores the importance of generating high-quality embeddings to ensure the inference process remains accurate and free from unintended biases. Poorly constructed embeddings can distort semantic relationships, leading to misleading conclusions and reinforcing existing biases within the data.

On the other side, the UMAP plot in Fig. 1 illustrates the embedding that best aligns with the ESCO taxonomy, as evaluated by the HSS metric introduced by Giabelli et al. [20]. To emphasize the effect of using a non-optimized word embedding model - trained with default parameters - Fig. 2 presents an alternative embedding generated from the same dataset, without validation against the ESCO taxonomy or the application of grid search optimization.

These algorithms learn dense vector representations by analyzing word co-occurrences within large text corpora. Unlike frequency-based models, static embeddings place semantically similar words closer in the vector space, allowing for more complex representations. However, a major limitation is that each word is assigned a single fixed vector, meaning polysemous words (e.g., "bank" as a financial institution vs. "bank" as a riverbank) cannot have distinct representations based on their meaning in different contexts. More recently, *contextual word embeddings* have emerged as a powerful alternative. These models, exemplified by transformer-based architectures such as BERT and GPT, generate word representations dynamically based on the surrounding context in a sentence. This enables them to distinguish between different meanings of the same word and more accurately capture linguistic nuances. Contextual embeddings have significantly improved performance in many NLP tasks, including text classification, machine translation, and question answering.

In this study, we use the FastText algorithm to train more than 100 models with different parameter combinations for each of 27 + 1 countries, representing over 4.5M OJAs.

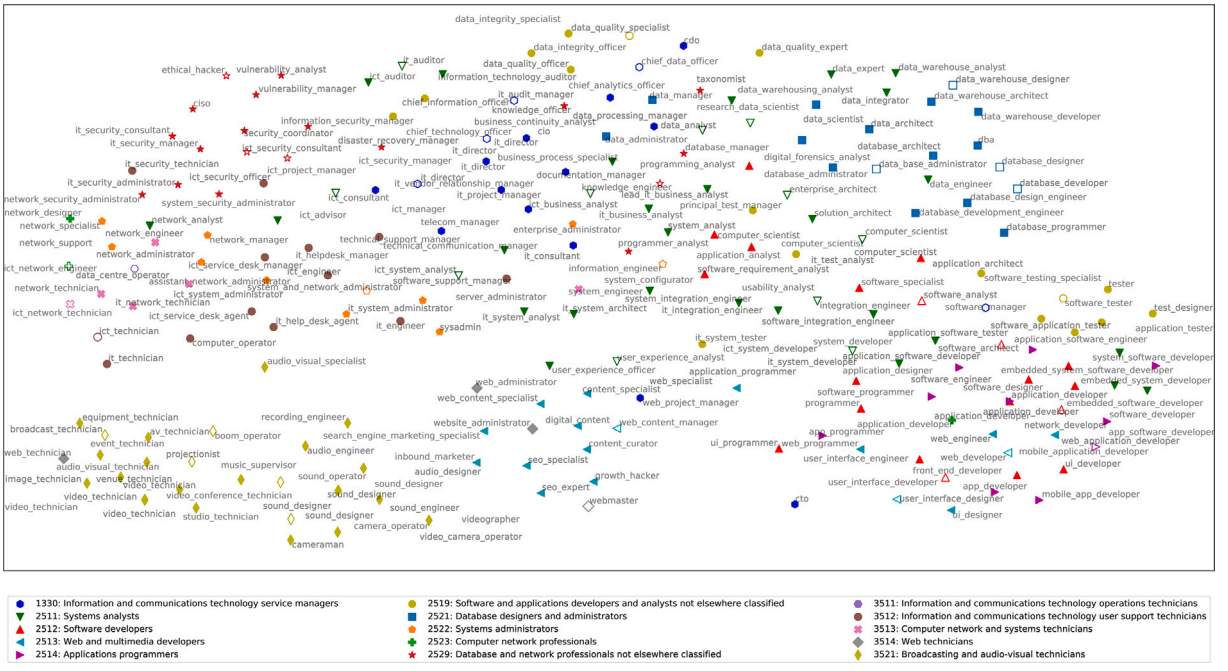


Fig. 1. UMAP plot of the **best** word embedding model, according to HSS metric [20]. The plot illustrates the ESCO concepts, and words belonging to each group are displayed, distinguishing between narrower occupations (empty shapes) and alternative labels (filled shapes). Trained over 2 million ICT-related jobs in the UK. Taken from [20].

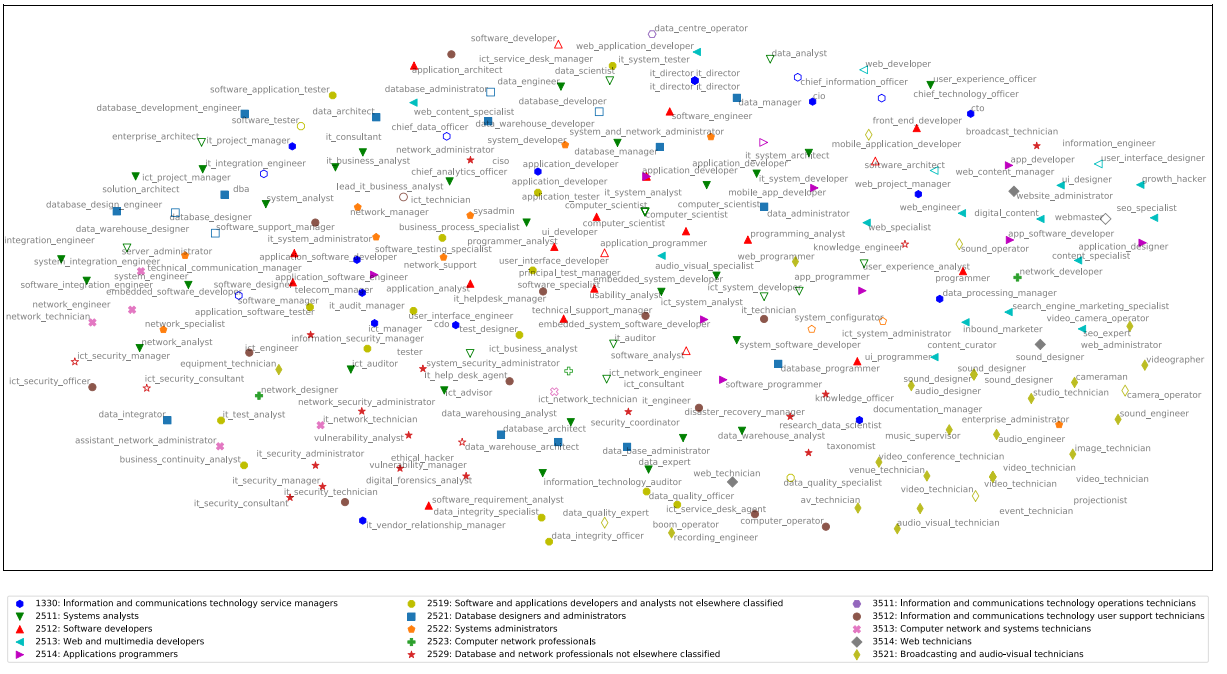


Fig. 2. UMAP plot of the *worst* word embedding model, according to HSS metric [20]. The plot shows the ESCO concepts and words belonging to each group, distinguishing between narrower occupations (shallow shape) and alternative labels (filled shape). Trained over 2 million ICT-related jobs in the UK.

Applications of word embeddings in the labor market. Machine learning techniques and word embeddings have been extensively used in the labor market for tasks such as job classification, skill extraction, and synthetic dataset generation. Online job postings offer a crucial resource for real-time labor market analysis, facilitating faster and more data-driven decision-making [8]. Research has shown that models trained on job vacancies can effectively address various challenges in this field.

Building on these resources, recent studies have also explored the structural organization of labor market data through graph-based systems. For instance, *GraphLMI* [18] utilizes graph databases to model the relationships between skills, occupations, and locations, providing a complementary perspective to purely vector-based representations for Labor Market Intelligence.

One notable application is job classification, where embeddings have been used to map job postings to standardized taxonomies. *JobBERT* by Decorte et al. [16] is a model fine-tuned on BERT with job-related texts to enhance classification performance. Similarly, Zhang et al. [47] leverages a multilingual pre-training model to enhance job classification across various languages, incorporating domain knowledge from job taxonomies. More recently, CareerBERT [40] has been proposed to map resumes and job descriptions into a shared embedding space based on the ESCO taxonomy to improve job recommendations. Despite these advancements, these models are often optimized for specific downstream tasks such as job recommendation or resumes matching. In contrast, our work addresses a different objective: providing a systematic framework to evaluate and select the most effective vector space models for labor market domain. By leveraging an intrinsic evaluation based on the ESCO taxonomy, we establish an objective criterion to assess how well different embeddings capture the formal semantic structure of occupations and skills. This approach ensures that the selected models are also methodologically aligned with official labor market standards across 28 languages.

Another critical task is skill extraction from job descriptions. Bhola et al. [5] introduces a deep learning approach to identify relevant skills associated with job postings, addressing the extreme multi-label nature of this problem. Recent advancements such as SkiLLMo [32] and Skillens [15] leverage Transformer models to achieve normalized skill extraction from job descriptions, directly mapping textual mentions to the ESCO taxonomy. Beyond classification and extraction, machine learning models have been employed to generate synthetic job postings and skill datasets. Clavié and Soulié [10] explores how large-scale language models can be used to match skills to job descriptions without requiring extensive labeled data. Additionally, Decorte et al. [17] proposes a training strategy for extracting skills from job descriptions, demonstrating how LLMs can enhance the availability of structured labor market data. These studies highlight the growing role of word embeddings and machine learning in enhancing labor market analytics, facilitating more accurate job classification, skill identification, and data augmentation for research and policy-making.

2.3. Taxonomic concepts similarity

A *taxonomy* is a structured classification system that organizes concepts into a hierarchical framework based on their relationships and levels of generality [19,20]. According to definition in [28], a taxonomy can be defined as:

Definition 1 (Taxonomy).
A *taxonomy* is defined as a couple $T = (C, H_C)$, where:

- C is the set of concepts $c \in C$ belonging to the domain of interest (i.e., the nodes of the taxonomy).
- H_C is a directed taxonomic binary relation between concepts, such that $H_C \subseteq \{(c_i, c_j) \mid (c_i, c_j) \in C^2, i \neq j\}$. This relation, denoted as $H_C(c_1, c_2)$, indicates that c_1 is a sub-concept of c_2, also known as the *IS-A* relation.

Taxonomies play a crucial role across domains, where they structure knowledge in ontologies and controlled vocabularies. In Natural Language Processing (NLP) and labor market analysis, taxonomies such as ESCO[6] (European Skills, Competences, and Occupations) provide a structured representation of skills and occupations, facilitating semantic interoperability across languages and regions.

In a taxonomic structure, concepts are typically arranged in a parent–child relationship, where broader categories encompass more specific subcategories. For example, in ESCO, the occupation *2512 - Software Developer* is a sub-concept of the broader concept *251 - Software and applications developers and analysts*, which in turn belongs to the more general concept of *2 - Professionals*. These hierarchical relationships can be leveraged to compute concept similarity, which is fundamental for tasks such as job matching, skills inference, and occupational classification.

Several measures are available to evaluate the similarity between concepts in a taxonomy. These measures can be broadly classified into two main categories: *path-based* and *Information Content (IC)-based* approaches. Path-based measures estimate similarity by analyzing the structure of the taxonomy, typically using the shortest path between two concepts or considering the depth of their lowest common ancestor, as proposed by Leacock et al. [26]. IC-based measures define similarity based on the probability of encountering a concept, assuming that more specific concepts carry higher information content. Approaches of this kind have been presented by Seco et al. [42].

2.4. The hierarchical semantic similarity at a glance

For our studies, we use the Hierarchical Semantic Similarity (HSS) developed by Giabelli et al. [20] as a similarity measure. The measure has then been implemented as a tool to perform taxonomy refinement [30]. The HSS is designed to measure the extent to which two concepts are related within a given taxonomy.

Intuitively, the metric is based on the concept of information content, which states that the lower the rank of a concept $c \in C$ that contains two entities, the higher the information content (IC) the two entities share. According to information theory, the IC of a concept c can be approximated by its negative log-likelihood: $IC(c) = -\log p(c)$ where $p(c)$ is the probability of encountering the

[6] https://esco.ec.europa.eu/en/classification.

concept c. Following Resnik [39], one can supplement the taxonomy with a probability measure $p : C \to [0, 1]$ such that for every concept $c \in C$, $p(c)$ is the probability of encountering an instance of the concept c. It follows that p (i) is monotonic and (ii) decreases with the rank of the taxonomy, i.e., if c_1 is a sub-concept of c_2, then $p(c_1) \leq p(c_2)$. This means the probability decreases as we move to more specific (deeper) concepts in the taxonomy. To estimate the values of p, Resnik employs the frequency of concepts within a large text corpus. However, we aim to infer similarity values within the semantic hierarchy to extend a taxonomy constructed by human experts. In this context, the Hierarchical Semantic Similarity measure is particularly useful, as it leverages the frequencies of concepts and entities within the taxonomy to compute p. Specifically, it estimates p as: $\hat{p}(c) = \frac{N_c}{N}$ where N is the cardinality, i.e., the number of entities (words), of the taxonomy and N_c is the sum of the cardinality of the concept c with the cardinality of all its hyponyms.

Note that $\hat{p}(c)$ is monotonic and increases with the size and generality of the concept c (i.e., N_c is larger for more general concepts), thus correctly reflecting the defined properties of p.

Given two words w_1 and w_2, Resnik defines $c_1 \in s(w_1)$ and $c_2 \in s(w_2)$ as all the concepts containing w_1 and w_2 respectively, i.e., the *senses* of w_1 and w_2. Therefore, there are $S_{w_1} \times S_{w_2}$ possible combinations of their word senses, where S_{w_1} and S_{w_2} are the cardinality of $s(w_1)$ and $s(w_2)$ respectively. Given that, Giabelli et al. [20] define $\mathcal{L}$ as the set of all the lowest common ancestor (LCA) for all the combinations of $c_1 \in s(w_1), c_2 \in s(w_2)$. The hierarchical semantic similarity between the words w_1 and w_2 can be defined as:

Definition 2 (Hierarchical Semantic Similarity (HSS)). The semantic similarity between two words w_1 and w_2 is computed as:

$$\texttt{HSS}(w_1, w_2) = \sum_{\ell \in \mathcal{L}} \hat{p}\left(\ell = L \mid w_1, w_2\right) \times IC(L) \tag{1}$$

where $\hat{p}\left(\ell = L \mid w_1, w_2\right)$ is the probability of ℓ being the Least Common Ancestor (LCA) of w_1 and w_2, computed using Bayes' theorem as:

$$\hat{p}\left(\ell = L \mid w_1, w_2\right) = \frac{\hat{p}\left(w_1, w_2 \mid \ell = L\right)\hat{p}(L)}{\hat{p}\left(w_1, w_2\right)} \tag{2}$$

Then, we define N_ℓ as the cardinality of ℓ and all its descendants. The numerator can be rewritten as:

$$\hat{p}\left(w_1, w_2 \mid \ell = L\right)\hat{p}(L) = \frac{S_{<w_1,w_2>\in\ell}}{|descend(\ell)|^2} \times \frac{N_\ell}{N} \tag{3}$$

where the first leg of the *rhs* is the class conditional probability of the pair $< w_1, w_2 >$ and the second one is the marginal probability of class ℓ. The term $|descend(\ell)|$ represents the number of subconcepts of ℓ. Since they could have at most one word sense w_i for each concept c, $|descend(\ell)|^2$ represents the maximum number of combinations of word senses $< w_1, w_2 >$ which have ℓ as LCA. $S_{<w_1,w_2>\in L}$ is the number of pairs of senses of words w_1 and w_2 which have L as LCA and $\frac{S_{<w_1,w_2>\in\ell}}{|\text{descend}(\ell)|^2}$ is the proportion of this maximum that is actually realized by the senses of w_1 and w_2. The denominator can be written as:

$$\hat{p}\left(w_1, w_2\right) = \sum_{k \in \mathcal{L}} \frac{S_{<w_1,w_2>\in k}}{|descend(k)|^2} \tag{4}$$

Benefits of using HSS. Unlike traditional measures that rely solely on lexical similarity or corpus-based co-occurrence, HSS explicitly considers the taxonomic distance between concepts, taking into account both their hierarchical depth and common ancestors. Given two occupations in a taxonomy, their HSS score reflects the degree to which they share commonalities: occupations that are direct siblings (i.e., sharing the same parent node) or have a close ancestor will exhibit a higher similarity than those that belong to distant branches of the hierarchy. By leveraging HSS, we can assess how well a word embedding model preserves the relationships defined in a taxonomy, providing an intrinsic evaluation of its effectiveness in capturing domain-specific knowledge. The following sections describe how we utilize this metric to compare different embedding models and identify those that best preserve the ESCO taxonomy relationships.

2.5. Embedding alignment

When word embeddings are trained independently on different corpora, their vector spaces are not directly comparable, even when derived from the same language. This misalignment is exacerbated in multilingual contexts, where embeddings are trained on corpora from different languages, making direct comparisons challenging, as exemplified in Fig. 3 from Ruder et al. [41]. Without alignment, embeddings from different models cannot be meaningfully compared, limiting their applicability in cross-lingual and cross-domain tasks. Aligning embedding spaces enables meaningful comparisons by mapping different vector spaces into a shared coordinate system. However, this process presents several challenges. First, structural differences arise due to variations in training data, corpus size, and linguistic properties, leading to inconsistencies between vector spaces. Second, selecting appropriate anchor points - words or concepts that serve as reference points for alignment - is critical, as an unsuitable choice can introduce distortions. Finally, the alignment method must balance computational efficiency and accuracy, ensuring that the transformed embeddings preserve semantic relationships while adapting effectively to the target space. Understanding the key challenges in cross-lingual language models is essential for evaluating how state-of-the-art methods address them. Mikolov et al. [36] observed that word vectors trained on monolingual data exhibit comparable topological structures across different languages. This observation laid the foundation for early alignment methods, which assumed that embedding spaces could be mapped through a simple linear transformation [36]. However, this assumption has

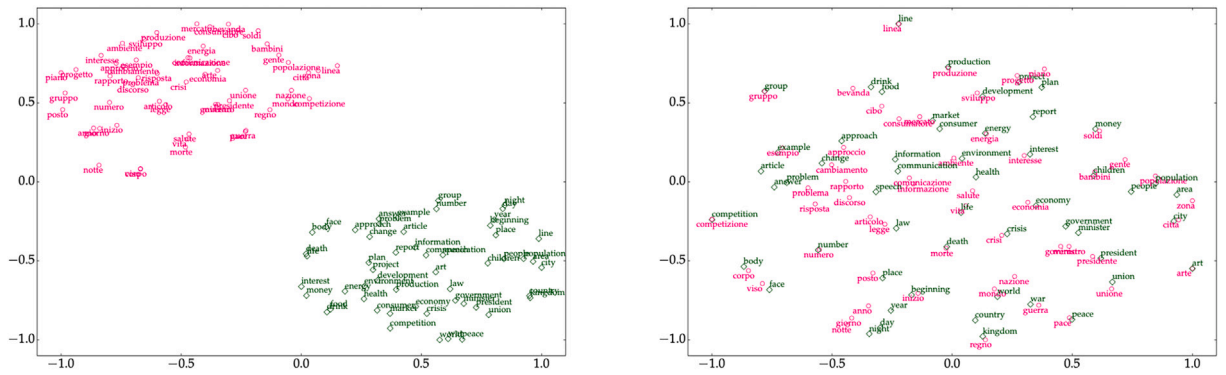


Fig. 3. An image from [41] depicting unaligned monolingual word embeddings (left) and word embeddings projected into a joint cross-lingual embedding space (right). Embeddings are visualized with t-SNE.

significant limitations. Linguistic variations in morphology, syntax, and semantics introduce complexities that challenge the validity of this hypothesis, making alignment more difficult in practice.

A key challenge in our study arises from our embeddings being trained separately on OJAs' corpora from different countries. Each dataset reflects unique linguistic and contextual nuances, leading to inherently non-comparable vector spaces. Even when the same concept appears across multiple countries, differences in terminology, language use, and data distributions cause variations in learned embeddings. Consequently, direct comparison of embeddings across countries is infeasible without an alignment mechanism. Various methods have been proposed to achieve this alignment; these approaches seek to enhance comparability while preserving the integrity of the original vector spaces. A common strategy for cross-lingual alignment involves constructing a seed lexicon - a collection of words with equivalent meanings in both corpora - serving as a reference for mapping embeddings. Within this research domain, several intuitive approaches have emerged. Artetxe et al. [1] introduced an unsupervised method that exploits the structural similarity of embedding spaces. Instead of relying on bilingual data, they use numerals as anchor points, assuming their meanings remain stable across languages. However, this approach can be sensitive to contextual variations in numeral usage across different corpora. Another notable contribution is the Hierarchical Cross-lingual Embedding Generation (HCEG) method by Azpiazu and Pera [4], which eliminates pivot language bias by leveraging linguistic hierarchies. Their approach enhances vocabulary induction, particularly for low-resource languages, though its computational complexity scales with the number of languages involved. A particularly influential method was proposed by Conneau et al. [14], who developed an unsupervised alignment technique that does not require parallel data. They learn a linear mapping between embedding spaces by combining adversarial training with Procrustean optimization. Their method also introduces a novel validation metric and the CSLS similarity measure, achieving results comparable to supervised approaches, even for distant and low-resource language pairs. Among these approaches, SeNSe [31] offers a distinct perspective. It defines a source model, representing the space to be aligned, and a target model, onto which the source model is mapped. Instead of relying on predefined seed lexicons, SeNSe selects optimal anchors dynamically by identifying words with stable meanings across corpora. It first builds a vocabulary of common words by translating terms and matching them across languages. A semantic similarity score (SNDCG) is then computed to assess alignment quality, retaining only high-scoring anchor pairs while removing duplicates. To ensure a balanced distribution, overly close anchors are filtered out. Finally, these selected anchors are used to learn an orthogonal transformation via Orthogonal Procrustes, preserving semantic relationships while mapping one vector space onto another. Alignment quality is then evaluated through tasks such as Bilingual Lexicon Induction (BLI).

We adopt SeNSe for two main reasons. First, the method has already been evaluated in multilingual scenarios involving Italian, Spanish, Finnish, and German aligned to English where it outperformed alternative state-of-the-art alignment approaches. This demonstrates its effectiveness in adapting multilingual vector spaces while preserving semantic consistency. Moreover, SeNSe has been previously applied in a labor market scenario, where embeddings trained on heterogeneous corpora were aligned and used to analyze occupational structures. This makes it well-suited to our cross-country analysis.

3. Building VEUCTOR on 4 million OJAs and 28 EU countries

Below, we present the workflow of embeddings generation and evaluation (Fig. 4) and alignment generation and evaluation (Fig. 5). Specifically, we introduce the framework for evaluating both the embedding models trained for each country and the models obtained through the alignment process. The goal is to establish a systematic methodology for assessing and identifying the best-performing models using extrinsic evaluation measures. This evaluation framework allows us to determine the most and least effective models based on the conducted assessment. The process consists of several key steps, which we describe in detail in the following sections.

Step 1 - embeddings pool generation. This step consists of two main phases: (i) preprocessing, where the raw text data from the OJA corpus is thoroughly cleaned, normalized, and prepared to ensure consistency across the dataset. The goal is to enhance the quality of the input data and make it suitable for further processing; (ii) training different models for each country, in which semantic

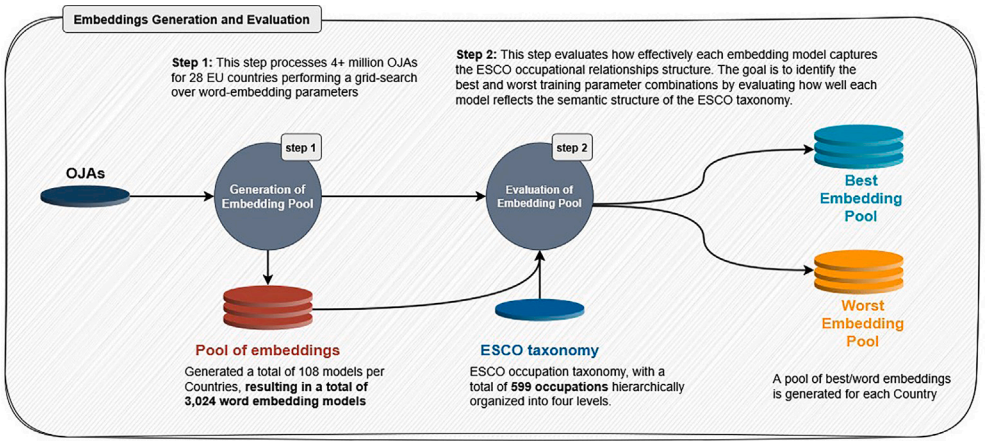


Fig. 4. Workflow for the generation and evaluation of word embedding models.

representations of words and phrases are derived from the preprocessed corpus. In this phase, FastText models are trained on the country-specific corpora using various combinations of training parameters. These embeddings serve as the semantic representations used in subsequent steps of the analysis.

Step 2 - embeddings pool evaluation. In this step, we evaluate the quality of the embeddings generated in Step 1, focusing on each country individually. The evaluation is extrinsic and relies on the ESCO taxonomy as an external resource to assess how effectively the embedding models capture occupational relationships. The goal is to identify the optimal and suboptimal training parameter combinations by evaluating how well each model captures the semantic structure of the ESCO taxonomy.[7] A high-quality embedding model should not only preserve the real-world relationships between occupations but also maintain the hierarchical organization defined by ESCO. Specifically, their vector representations in the embedding space should exhibit high similarity if two occupations are closely related in ESCO - whether as siblings (i.e., sharing the same parent category) or in a direct parent–child relationship. Conversely, occupations that are distant within the hierarchy should show lower similarity. To assess the quality of the embeddings, we employed two key metrics: cosine similarity and Hierarchical Skill Similarity (HSS), the latter of which quantifies the semantic relationship between occupations in the ESCO taxonomy [29]. For every pair of occupations (ISCO 4-digit), we first calculate the cosine similarity between their embedding vectors. We then compute the HSS values for the same set of occupation pairs. Spearman's rank correlation coefficient (ρ) is applied to quantify the relationship between the cosine similarity scores and the HSS values. A higher correlation indicates a stronger alignment between the occupational concepts captured by the embedding models and the relationships defined within ESCO. Algorithm 1 in the Appendix outlines the execution of Steps 1 and 2: for each country and each combination of training parameters, the model is trained and evaluated by computing the Spearman rank correlation coefficient between the cosine similarity scores and the HSS values. At the end of the process, the best and worst parameter sets are determined for each country.

Step 3 - alignment embeddings pool generation. In this step, we align the embedding models using the SeNSe technique, as described in the previous section. The goal is to obtain comparable vector models across countries to enable inter-country analysis. The main challenge is that these models were trained on different corpora and with different hyperparameter settings, making them inherently non-comparable. This discrepancy arises because the training process was not consistent across countries. To address this issue, we first select a common set of hyperparameters to standardize the models. Specifically, we choose the best-performing hyperparameter set based on the evaluation from the previous steps. This ensures that the 28 models (one per country) being aligned all share the same hyperparameter configuration, facilitating a more reliable comparison. Using an alignment technique, we transform the original vector spaces into a new set of aligned spaces, making them comparable. The SeNSe approach aligns source vector spaces to a predefined target space. In our case, we designate the UK model as the target space, meaning that all other country-specific vector spaces are aligned with the UK model. The alignment is performed pairwise, where each country's vector space acts as the source, and the UK model serves as the target. By aligning all 27 other countries to a common target, we ensure that any two countries can be directly compared within the same aligned space. The alignment algorithm requires tuning several parameters. In the previous step, we generated multiple alignment variations for different parameter combinations for each country, allowing us to assess the impact of different alignment configurations on model alignment.

Step 4 - alignment embeddings pool evaluation. In this step, we evaluate the aligned models for each country, which were generated in the previous stage. The evaluation criterion used is a cross-lingual semantic fitting score (*CLS score*): *Cross-Lingual* highlights

[7] Our use of ESCO does not imply an endorsement of its superiority or optimality. Our approach is rather practical. ESCO is the official EU taxonomy and has become the de facto standard taxonomy used in Europe, making it the most logical and operationally viable benchmark for our purposes.

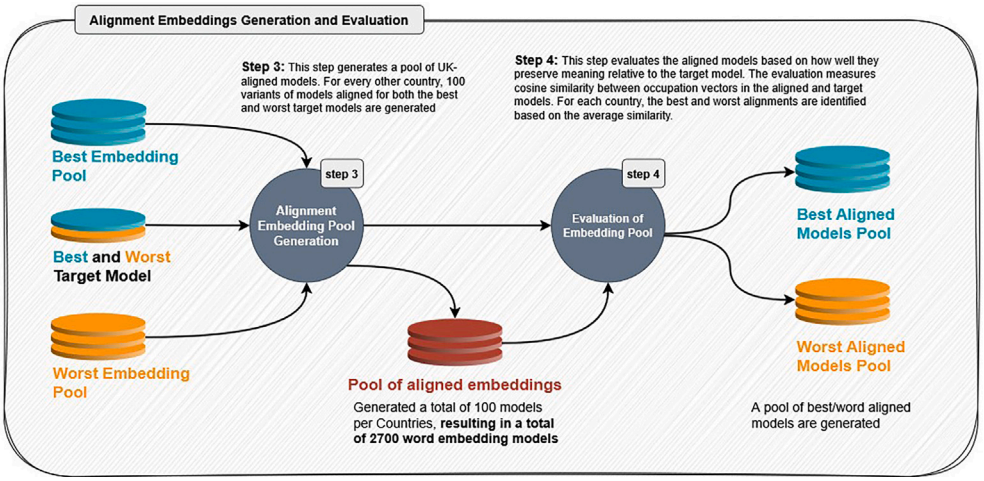


Fig. 5. Workflow for the generation and evaluation of aligned models.

that the comparison is performed between different languages, *Semantic* emphasizes that the evaluation focuses on preserving the meaning in the aligned vectors and *Fitting score* indicates that we measure how well the aligned model adapts to the target space. The evaluation assesses the aligned model based on its ability to generate similar vectors for the same concept when expressed in both the source and target models. For this assessment, we consider occupations that appear in the vocabularies of both the source and target models. For each occupation, we compute the cosine similarity between the corresponding vectors in the aligned and target models. A higher cosine similarity indicates a stronger alignment between the concept in the source language and its counterpart in the target language. For each country, the best-aligned model is identified as the one that maximizes the average cosine similarity across all occupations. Similarly, the worst-aligned model is determined to have the lowest average cosine similarity. Algorithm 2 in the Appendix describes the process for steps 3 and 4. The input consists of the set of parameters used for alignment, the source models $\{\mathcal{M}_c\}c \in C$ (where each model represents the best-performing model for a specific country c), and the list of occupations used for evaluation. For each country and each parameter combination, a new aligned model ($\mathcal{A}_{c,\theta}$) is generated. The cosine similarity is then computed between the vectors corresponding to the same occupation in the target and aligned models. The average cosine similarity across all occupations is calculated for each model, and the best/worst parameter combination for country c is selected as the one that maximizes/minimizes the similarity score ($score_{\mathcal{A}_{c,\theta}}$) among the different aligned models for c.

4. Experimental results

The ESCO taxonomy. ESCO (European Skills, Competences, Qualifications and Occupations) is the European classification of skills, competences, and occupations. It provides a multilingual dictionary of occupations and skill requirements organised along two main pillars. The first is the Occupation pillar, which is referenced to the ISCO08 standard. The second is the Skills pillar, which lists and describes competencies/skills that are linked to occupations. Therefore, ESCO provides a list of occupations and related skills, organised as a network; nonetheless, it gives no information on the importance of skills in the considered occupation.

OJA data. In the online job market, a job advertisement is a document containing two main text fields: a title and a full description. The title briefly summarizes the job position, while the full description field typically includes the position details and the relevant skills the employee must possess (see, e.g., [34]). The OJAs used here have been collected as part of the Web Intelligence Hub (WIH), which is a component of Eurostat's Trusted Smart Statistics (TSS) initiative, aiming to leverage web data for statistical purposes through the analysis of new data sources using advanced technologies such as artificial intelligence to integrate traditional sources used in official statistics. The WIH-OJA use case, developed by Cedefop and Eurostat and added to the WIH in 2021, is devoted to collecting and processing online job advertisements from sources in 32 countries (EU, EEA and the UK).

The OJA-WIH representative sample. Within the WIH initiative, Eurostat developed a representative sample of the entire OJA dataset, which is now composed of over 450 million unique OJAs collected since 2019. The sample aims to allow the community to work on a smaller dataset while preserving the distributional characteristics of online job ads. Specifically, the WIH-OJA-NLPv1 table seeks to leverage the richness of information extracted from online OJA portals. The goal of Natural Language Processing (NLP) data flows is to utilize the raw job titles and descriptions collected via web scraping techniques. The NLP samples serve two objectives, system development – a reference set of observations is necessary to test the impact of new techniques on the classification results – and research distribution – to enable research initiatives on OJA data. The sample is stratified by language, with the exception of very small ones and seven of the variables under which the OJAs are classified. The sampling is balanced and covers all possible values of the classification variables. The stratification variables include occupation (ISCO-08 III digit level), type of contract (permanent,

Table 1
Overview of OJAs statistics per country and year, including linguistic and structural features of the datasets.

Country	Country Code	Number of OJAs			Languages	Occupations	Vocabulary Size (# of words)	Text Lengths		
		2020	2021	Total				Avg. Title	Avg. Desc.	Avg. Tokens
Austria	AT	43,747	93,057	136,804	de	415	1,005,288	4.65	213.97	62.46
Belgium	BE	59,267	288,286	347,553	de, fr, nl	423	2,148,097	4.33	245.63	66.19
Bulgaria	BG	11,798	54,125	65,923	bg	525	670,812	9.14	281.95	74.86
Cyprus	CY	1,831	8,657	10,488	el	348	123,220	5.41	229.00	82.10
Czechia	CZ	7,223	56,598	63,821	cs	394	638,677	6.38	282.28	84.32
Germany	DE	96,779	391,060	487,839	de	492	2,955,779	5.63	246.48	62.92
Denmark	DK	8,302	38,625	46,927	da	369	728,429	6.81	470.79	146.84
Estonia	EE	1,919	13,742	15,661	et	337	157,323	3.65	142.46	53.85
Greece	EL	7,441	35,118	42,559	el	402	420,798	4.85	226.45	75.25
Spain	ES	43,003	192,681	235,684	es	419	1,170,647	5.22	233.57	58.12
Finland	FI	6,993	33,467	40,460	fi	395	403,520	3.386	192.604	74.603
France	FR	117,872	546,336	664,208	fr	510	2,865,121	5.651	294.121	70.191
Croatia	HR	6,093	30,874	36,967	hr	399	336,653	4.857	295.936	83.535
Hungary	HU	8,901	72,172	81,073	hu	403	796,307	4.509	227.544	72.518
Ireland	IE	26,698	94,760	121,458	en	409	720,802	4.173	304.217	85.605
Italy	IT	109,298	253,133	362,431	it	422	1,536,694	4.741	208.025	52.705
Lithuania	LT	5,279	25,229	30,508	lt	388	218,319	4.995	174.565	58.232
Luxembourg	LU	4,415	18,069	22,484	de, fr	326	193,320	5.854	274.225	76.536
Latvia	LV	4,950	21,201	26,151	lv	374	234,378	5.024	247.562	78.94
Malta	MT	1,176	5,925	7,101	mt	285	63,318	3.528	272.847	90.594
Netherlands	NL	60,613	316,420	377,033	nl	423	2,744,164	3.908	395.328	101.918
Poland	PL	48,400	122,337	170,737	pl	419	1,453,477	4.923	371.008	99.978
Portugal	PT	35,172	164,372	199,544	pt	449	1,679,525	5.274	307.137	80.549
Romania	RO	14,546	71,312	85,858	ro	411	612,775	4.854	231.257	67.846
Sweden	SE	28,410	136,420	164,830	sv	423	1,263,516	4.893	392.263	95.109
Slovenia	SI	2,495	11,221	13,716	sl	358	117,635	5.177	147.382	51.058
Slovakia	SK	8,349	55,080	63,429	sk	378	565,275	4.784	298.41	86.376
United Kingdom	UK	185,333	502,930	688,263	en	492	2,802,098	4.207	313.692	85.55
Total	28	921,738	3,497,491	4,419,229	23					

temporary, apprenticeship or traineeship, and self-employed), salary, working time, education level, economic activity (NACE divisions), and experience. The v1 samples contain OJAs stratified by all classified variables, and each stratum has a maximum of 50 observations, resulting in a total number of observations of 4,610,821. Considering the size of the sample and the level of detail provided, the database can be fruitfully employed to conduct research beyond the scope of quality. In this study, we use the NLP sample v1, release r20221217.

4.1. Data pre-processing

In this step, raw text data undergoes several transformations to ensure consistency and remove noise that may affect subsequent analysis. We form the target texts by concatenating the job title with its description, as the title often contains relevant information such as the occupation or the required skills. First, unnecessary elements such as HTML tags, special characters, and URLs are removed, followed by the stripping of numerical values, certain symbols, and punctuation to retain only alphabetic content. The text is then normalized by converting all characters to lowercase, preventing inconsistencies due to case sensitivity. Additionally, sequences of characters representing meaningful multi-word expressions (n-grams) are identified[8] and replaced to preserve important phrases as single units. The preprocessing was further tailored to the corpus of each country, with linguistic operations adapted to the official languages of each (e.g., French, Dutch, and German for Belgium). Language-specific stopwords were excluded to refine the content, retaining only the most relevant terms. These steps collectively produce a clean and uniform dataset, improving its quality while enabling subsequent fastText models to better capture the corpus's concepts, identify linguistic relationships, and enhance advanced applications such as content analysis and creating rich semantic representations.

In addition to the preprocessing steps described above, Table 1 provides an overview of the OJAs datasets, detailing key statistics per country and year. The dataset covers 28 countries and includes information on the number of OJAs for 2020 and 2021, their total count, and the official languages used in each country. The table also highlights additional structural characteristics, including the number of distinct occupations, vocabulary size, and the average lengths of titles, descriptions, and tokens. These features underline the dataset's heterogeneity and underscore the need for tailored preprocessing to extract meaningful patterns and enhance the performance of subsequent word embedding models in capturing semantic relationships and domain-specific concepts effectively.

[8] https://radimrehurek.com/gensim/models/phrases.html.

4.2. Embeddings pool generation

In this section, we describe the process of generating, evaluating, and aligning embeddings for various occupations across different countries. The method produces occupational embeddings and evaluates their similarities using statistical metrics. The key steps of this process are outlined below.

In this stage, we train FastText models on job advertisement data from 28 countries to generate a diverse pool of embeddings. For each country, a separate corpus of job advertisements is processed in order to reflect linguistic and contextual differences. We therefore explore how different training choices affect embedding quality by performing a grid search over a set of FastText hyperparameters, including:

Embedding size (v_{size}): The dimensionality of the word vectors. Smaller vectors may lead to underfitting, where the embeddings are too simplistic to capture fine-grained semantic details, whereas larger vectors tend to capture more complex relationships but may require more data and computational resources.

Number of epochs (τ): The number of epochs was tested to achieve a balance between training time and embedding quality. Lower values allow for faster training but may result in underfitting, where the model fails to capture the full complexity of the data. Higher values provide the model with more opportunities to refine the embeddings and capture more nuanced semantic relationships, but they also increase the risk of overfitting.

Algorithm (A): This parameter defines the architecture of the neural network used during training. The Skip-Gram (SG) algorithm predicts context words from a given target word. It is particularly effective for smaller datasets, as it is better at capturing rare words and their semantic relationships. The Continuous Bag of Words (CBOW) algorithm predicts a target word based on its surrounding context. CBOW is generally more efficient for larger corpora, as it aggregates context words to make predictions, resulting in faster training times and improved performance on larger datasets.

Hierarchical softmax (hs): This parameter determines whether hierarchical softmax is used, which can improve training efficiency when dealing with large vocabularies.

Learning rate (α): This parameter controls the speed at which the model updates its weights during training. A lower learning rate results in slower, more gradual updates, which can help the model converge more precisely; however, it may require more epochs to reach optimal performance. A higher learning rate enables faster updates, which can accelerate training but may lead to overshooting the optimal solution or instability in the learning process. The choice of learning rate has a significant impact on the balance between training time and the quality of the final embeddings.

For each country, 108 models were trained by varying the parameters across the following ranges: $v_{size} \in \{50, 100, 300\} \times \tau \in \{10, 50, 100\} \times \mathbf{A} \in \{SG, CBOW\} \times hs \in \{0, 1\} \times \alpha \in \{0.01, 0.05, 0.1\}$. This was an exhaustive grid search without early stopping; each of the 108 parameter combinations per country was trained to completion to ensure a comprehensive evaluation.

4.3. Embeddings pool evaluation

This section describes the evaluation of the model pool trained for each country. The best and worst training parameter combinations were identified for each model and country using the Spearman rank correlation coefficient between the cosine similarity scores and HSS values for each occupation pair. Table 2 presents the evaluation results, showing the values corresponding to the training parameters used.

What emerges is that no single parameter combination consistently yields the best results across countries, as performance depends on both the chosen parameters and the training corpus. For example, for the selected algorithm, Skip-gram (SG) performs best in most cases (18 countries), whereas Continuous Bag of Words (CBOW) performs worst (16 countries). However, it is also observed that in some cases (12 countries), the algorithm is both the best and the worst, depending on the specific context.

It is also evident that the Spearman correlation values (ρ) are generally very low, with some of the worst cases even showing negative values. This indicates a significant task challenge, highlighting the difficulty of all models in accurately capturing the hierarchical relationships between occupations as defined in the ESCO taxonomy. Such low or negative correlation values may also reflect a gap between the real-world labor market and the structure of ESCO, suggesting that the embedding models struggle to fully reflect the semantic relationships inherent in the ESCO taxonomy.

4.3.1. Comparison with pretrained LLMs

A key question in the embedding-based analysis is whether a domain-specific model, trained directly on the OJAs corpus, can outperform large-scale pre-trained language models (LLMs) in capturing occupational similarities. While LLMs benefit from extensive training on diverse and heterogeneous corpora, their ability to accurately preserve structured relationships between occupations remains uncertain. To investigate this, we compare the best UK-trained embedding model against four pre-trained models, evaluating their performance using the same methodology adopted for pool evaluation.

To ensure a consistent and equitable comparison, we utilize these Transformer-based LLMs as embedding models in an encoder-only configuration. This approach aligns with our FastText-based methodology, as both architectures are tasked with encoding skill and occupation concepts into a fixed-length vector space for semantic analysis.

To select appropriate pre-trained models for comparison, we relied on the benchmark constructed by Muennighoff et al. [37], which provides a comprehensive evaluation of text embedding models. Their framework, the Massive Text Embedding Benchmark

Table 2
Best and worst FastText parameter combinations for each country, along with the corresponding Spearman correlation index (ρ) and its p-value (p_ρ). Training times of model pools for different countries are also given.

	Best Training Parameters							Worst Training Parameters							Machine Training Time (108 models per country)			
country	v_{size}	τ	A	hs	α	ρ	p_ρ	v_{size}	τ	A	hs	α	ρ	p_ρ	Avg.	Min	Max	Total Time
AT	300	10	SG	✓	0.1	0.088	2.481E-08	300	50	SG	✓	0.05	-0.069	1.034E-05	1h 40m	0h 10m	6h 35m	7d 12h
BE	300	100	SG	✓	0.1	0.247	1.652E-70	50	50	SG	✓	0.1	0.053	2.037E-04	1h 45m	0h 10m	6h 15m	7d 22h
BG	100	50	SG	✗	0.05	0.070	4.307E-04	300	100	CBOW	✓	0.1	-0.176	3.293E-20	0h 14m	0h 01m	0h 56m	1d 02h
CY	300	10	SG	✓	0.01	0.252	5.722E-81	50	100	CBOW	✗	0.1	0.001	9.149E-01	0h 13m	0h 01m	0h 56m	1d 00h
CZ	100	50	SG	✗	0.01	0.150	6.430E-14	100	50	CBOW	✓	0.1	0.005	7.877E-01	0h 18m	0h 01m	1h 11m	1d 09h
DE	100	100	SG	✓	0.1	0.143	8.582E-23	300	50	SG	✗	0.1	-0.060	4.624E-05	3h 17m	0h 17m	12h 02m	14d 19h
DK	300	50	CBOW	✓	0.01	0.256	9.118E-62	50	50	SG	✓	0.1	0.087	3.264E-08	1h 08m	0h 06m	4h 36m	5d 02h
EE	300	50	SG	✗	0.01	0.084	1.516E-10	100	100	CBOW	✓	0.05	-0.105	1.366E-15	0h 16m	0h 01m	1h 08m	1d 05h
EL	300	50	SG	✗	0.1	0.381	5.528E-167	300	100	CBOW	✗	0.1	0.150	9.620E-26	0h 23m	0h 02m	1h 37m	1d 18h
ES	300	10	SG	✗	0.05	0.335	6.348E-46	300	10	SG	✓	0.1	0.054	2.368E-02	1h 43m	0h 09m	7h 05m	7d 18h
FI	300	10	CBOW	✓	0.1	0.167	1.338E-28	50	100	CBOW	✓	0.1	-0.044	3.937E-03	0h 23m	0h 02m	1h 37m	1d 19h
FR	100	10	CBOW	✓	0.05	0.342	7.193E-42	50	10	CBOW	✗	0.01	0.170	4.893E-11	3h 21m	0h 17m	12h 32m	15d 02h
HR	300	10	CBOW	✗	0.01	0.285	4.971E-53	100	50	CBOW	✓	0.1	0.038	4.660E-02	0h 39m	0h 03m	2h 45m	2d 23h
HU	50	50	SG	✗	0.01	0.117	6.176E-11	300	100	CBOW	✗	0.1	-0.061	6.085E-04	0h 20m	0h 01m	1h 15m	1d 12h
IE	300	10	SG	✓	0.1	0.230	1.317E-56	50	10	SG	✓	0.01	-0.018	2.134E-01	1h 12m	0h 06m	5h 00m	5d 11h
IT	50	10	CBOW	✗	0.01	0.189	4.580E-27	100	10	CBOW	✓	0.1	-0.068	1.417E-04	1h 46m	0h 09m	6h 25m	7d 23h
LT	100	10	SG	✗	0.05	0.210	2.798E-28	50	100	CBOW	✓	0.05	-0.039	3.628E-02	0h 24m	0h 01m	1h 52m	1d 20h
LU	50	100	CBOW	✓	0.1	0.123	4.869E-12	100	10	SG	✓	0.1	-0.078	8.800E-06	0h 26m	0h 03m	1h 52m	1d 23h
LV	100	50	SG	✗	0.05	0.291	8.649E-43	50	50	CBOW	✓	0.1	0.011	6.094E-01	0h 28m	0h 02m	1h 58m	2d 03h
MT	300	50	SG	✓	0.1	0.338	5.270E-138	50	50	CBOW	✓	0.1	0.047	6.633E-04	0h 09m	0h 00m	0h 41m	0d 16h
NL	100	100	CBOW	✗	0.1	0.246	1.369E-41	100	100	CBOW	✓	0.01	0.121	3.405E-11	2h 13m	0h 13m	7h 28m	9d 23h
PL	50	10	CBOW	✗	0.01	0.291	1.181E-44	50	50	SG	✓	0.1	0.026	2.284E-01	1h 15m	0h 05m	4h 55m	5d 15h
PT	50	10	CBOW	✗	0.01	0.324	2.508E-64	300	100	SG	✗	0.1	0.053	7.035E-03	1h 05m	0h 07m	4h 44m	4d 21h
RO	300	100	SG	✗	0.01	0.220	2.204E-40	50	100	CBOW	✓	0.1	0.058	4.923E-04	0h 53m	0h 04m	3h 41m	4d 00h
SE	300	100	SG	✗	0.1	0.083	4.632E-08	100	10	SG	✓	0.01	-0.105	1.160E-12	0h 53m	0h 04m	3h 18m	4d 00h
SI	50	50	SG	✗	0.05	0.219	1.124E-22	50	50	CBOW	✓	0.05	0.013	5.677E-01	0h 21m	0h 02m	1h 34m	1d 13h
SK	300	10	SG	✗	0.05	0.170	3.319E-13	50	100	SG	✓	0.1	-0.017	4.775E-01	0h 29m	0h 01m	2h 29m	2d 05h
UK	50	50	CBOW	✗	0.01	0.269	1.017E-55	300	50	SG	✗	0.1	0.071	2.290E-05	3h 50m	0h 20m	14h 16m	17d 05h

(MTEB), assesses models across various embedding tasks, offering insights into their relative strengths and weaknesses. MTEB evaluates models based on factors such as computational efficiency, embedding dimensionality, and performance on multiple NLP tasks. For our selection, we focused on the Semantic Textual Similarity (STS) task, as it closely aligns with our own evaluation methodology for assessing occupational similarity. The continuously updated leaderboard, ranking the best-performing models, is publicly available on Hugging Face platform.[9] According to the leaderboard, we selected 4 open-source pre-trained models: *BAAI/bge-large-en-v1.5*[10] [46] with 335M params, *thenlper/gte-large*[11] [27] with 335M params, *Lajavaness/bilingual-embedding-base*[12] [43] with 278M params and *intfloat/multilingual-e5-large-instruct*[13] [45] with 7.11B params. These models represent the state-of-the-art in Transformer-based architectures optimized for embedding tasks, primarily leveraging BERT-like (e.g., mBERT or XLM-R) encoder structures to generate high-quality contextual representations.

To ensure a comprehensive and fair comparison, we evaluated each pre-trained model in three distinct settings:

1. **Zero-Shot:** Using the models out-of-the-box without any further training. This tests their inherent, general-purpose knowledge of occupational semantics.
2. **Fine-Tuned on ESCO:** We fine-tuned the models on the text descriptions of skills and occupations within the ESCO taxonomy. This tests their ability to adapt to the structure of our target benchmark directly, providing them with explicit knowledge of the hierarchical relationships we aim to capture.
3. **Fine-Tuned on UK OJA Dataset:** We fine-tuned the models on the same corpus of UK job advertisements used to train our FastText models. This evaluates their performance when granted domain-specific adaptation on identical data, allowing for a direct and equitable comparison of modelling architectures by controlling for the training corpus.

The results are summarized in Table 3. The UK-trained model achieved the highest correlation score of 0.269, surpassing all four pre-trained models in terms of performance. Additionally, the p-values for all correlations are effectively zero, confirming the statistical significance of these results.

This demonstrates that even when LLMs are provided with the same domain-specific data (OJA fine-tuning) or explicit taxonomic knowledge (ESCO fine-tuning), our simpler, dedicated embedding approach better captures the hierarchical relationships defined in the ESCO taxonomy for this task. The advantage can be attributed to the fact that the FastText model was designed and trained

[9] https://huggingface.co/spaces/mteb/leaderboard.
[10] https://huggingface.co/BAAI/bge-large-en-v1.5.
[11] https://huggingface.co/thenlper/gte-large.
[12] https://huggingface.co/Lajavaness/bilingual-embedding-base.
[13] https://huggingface.co/intfloat/multilingual-e5-large-instruct.

Table 3
Comparison of pre-trained LLMs under different adaptation strategies. For each configuration (Zero-Shot, FT on ESCO, FT on OJA), the highest Spearman correlation (ρ) is underlined. The overall best result is highlighted in bold.

	LLM open-models	ρ	p_ρ
Zero-Shot	bge-large-en-v1.5	0.183	2.773E-28
	gte-large	0.193	2.207E-31
	bilingual-embedding-base	0.214	4.332E-38
	multilingual-e5-large-instruct	<u>0.246</u>	2.834E-50
Fine-Tuning on ESCO	bge-large-en-v1.5	0.178	1.088E-26
	gte-large	0.175	5.837E-26
	bilingual-embedding-base	0.198	5.255E-33
	multilingual-e5-large-instruct	<u>0.217</u>	2.042E-39
Fine-tuning on NLP UK dataset	bge-large-en-v1.5	0.239	7.132E-95
	gte-large	0.211	1.861E-74
	bilingual-embedding-base	0.220	6.411E-81
	multilingual-e5-large-instruct	<u>0.227</u>	4.558E-86
	best UK-trained model (our)	**0.269**	1.017E-55

directly to optimize geometric relationships in a vector space, which aligns well with the semantic similarity assessment performed by the HSS metric. Despite having orders of magnitude fewer parameters, its focused training objective proves more effective for this specific application than the more general-purpose representations of LLMs.

4.4. Alignment embeddings pool generation

In this step, pools of embedding models aligned to the UK model are generated for each country. To align models using SeNSe, they must be trained with the same set of hyperparameters. As previously described, SeNSe aligns vector spaces by applying matrix transformations to shift the source space toward the target space, allowing the generation of new vectors for terms in the source space. For example, if the vector dimensions differ, this transformation cannot be applied, making prior standardization of hyperparameters essential. The best parameter combination across all countries was identified based on the highest average Spearman correlation. The optimal configuration consists of the following hyperparameters: $v_{size} = 300$, $\tau = 50$, $\mathbf{A} = \mathrm{SG}$, $hs = 0$, $\alpha = 0.01$. Therefore, only the 28 models trained with this configuration were considered for alignment. The UK model was chosen as the target for the alignment process, and all other country-specific models were aligned to it pairwise. This approach ensures that all models are mapped to a common space, enabling direct comparison across countries. SeNSe relies on several alignment parameters to optimize the transformation process. One of the most critical is the maximum allowed NDCG value for selecting the best anchors. The choice of this parameter significantly affects the quality of the alignment: lower thresholds result in fewer selected anchor points, meaning only the most confident semantic correspondences are used. While this can lead to a more stable transformation, it may also exclude useful but less obvious alignments. Higher thresholds increase the number of selected anchor points, leading to a denser alignment. This can help capture a broader range of semantic relationships but might introduce noise if less reliable correspondences are included. Given the potential impact of this parameter on alignment quality, it is crucial to test different values systematically. For this study, threshold values ranging from 0 to 0.99 were tested in increments of 0.01, yielding 100 aligned models per country. This exhaustive approach ensures that we can identify the optimal alignment configuration for each country, balancing alignment accuracy and semantic consistency. Selecting the best-performing alignment model is essential, as a poor alignment could distort cross-lingual analyses, leading to misleading conclusions.

4.5. Alignment embeddings pool evaluation

After generating multiple aligned models, as described in the previous section, their quality was assessed using the Cross-Lingual Semantic Fitting Score (CLS score). This evaluation follows the procedure outlined in Algorithm 2 in the Appendix. It aims to determine how well each aligned model preserves the semantic relationships of occupations relative to the target model. The evaluation process consists of the following steps: (i) the occupations present in both the source and target model vocabularies are considered, (ii) for each matched occupation, the cosine similarity between its vector representation in the aligned model and in the target model is computed and (iii) the unweighted average cosine similarity across all occupations serves as the final CLS score, indicating the overall alignment quality. Table 4 presents the evaluation results for each country's best and worst alignments. The table reports the NDCG threshold used to select anchor points during alignment, the CLS score, which measures semantic consistency, and the confidence interval for the mean CLS score.

The results show a clear trend: the best alignments tend to have a lower NDCG threshold, meaning that a larger number of anchor points were used. This results in a more robust alignment, as more semantic relationships between the source and target spaces are preserved. Conversely, the worst alignments correspond to higher NDCG thresholds, which impose a stricter selection on anchor points. The few available anchors lead to a weaker alignment. These findings align with the work in [31], confirming that many anchor sets generally lead to better alignment performance.

Table 4
Best and worst NDCG thresholds for each country, along with the corresponding *CLS score*.

	Best Alignment Model			Worst Alignment Model		
country	NDCG threshold	CLS score	95% c.i	NDCG threshold	CLS score	95% c.i
AT	0.01	0.77	0.766 - 0.774	0.99	0.479	0.469 - 0.489
BE	0.0	0.809	0.805 - 0.814	0.99	0.428	0.418 - 0.438
BG	0.02	0.757	0.752 - 0.762	0.99	0.396	0.385 - 0.407
CY	0.03	0.7	0.695 - 0.705	0.99	0.444	0.433 - 0.455
CZ	0.02	0.749	0.743 - 0.756	0.99	0.441	0.427 - 0.455
DE	0.0	0.765	0.761 - 0.769	0.99	0.479	0.468 - 0.489
DK	0.0	0.761	0.755 - 0.768	0.99	0.41	0.395 - 0.424
EE	0.01	0.66	0.655 - 0.666	0.99	0.451	0.44 - 0.462
EL	0.02	0.774	0.769 - 0.778	0.99	0.442	0.432 - 0.451
ES	0.01	0.785	0.781 - 0.789	0.99	0.474	0.465 - 0.484
FI	0.01	0.727	0.722 - 0.732	0.99	0.446	0.436 - 0.456
FR	0.0	0.77	0.764 - 0.776	0.99	0.417	0.401 - 0.432
HR	0.02	0.727	0.72 - 0.733	0.99	0.454	0.442 - 0.467
HU	0.03	0.765	0.758 - 0.771	0.99	0.411	0.398 - 0.424
IE	0.26	0.835	0.83 - 0.84	0.99	0.392	0.377 - 0.406
IT	0.01	0.787	0.781 - 0.793	0.99	0.463	0.448 - 0.477
LT	0.01	0.725	0.718 - 0.732	0.99	0.485	0.47 - 0.5
LU	0.03	0.746	0.741 - 0.751	0.99	0.407	0.396 - 0.419
LV	0.01	0.696	0.688 - 0.703	0.99	0.478	0.465 - 0.491
MT	0.1	0.739	0.732 - 0.746	0.99	0.41	0.393 - 0.427
NL	0.0	0.798	0.792 - 0.804	0.99	0.429	0.416 - 0.443
PL	0.0	0.768	0.761 - 0.775	0.99	0.467	0.452 - 0.483
PT	0.02	0.812	0.806 - 0.817	0.99	0.443	0.428 - 0.458
RO	0.11	0.784	0.778 - 0.79	0.99	0.385	0.37 - 0.401
SE	0.0	0.778	0.772 - 0.784	0.99	0.417	0.404 - 0.429
SI	0.0	0.627	0.619 - 0.634	0.99	0.445	0.434 - 0.457
SK	0.03	0.734	0.729 - 0.738	0.99	0.445	0.435 - 0.455

A notable exception is observed for Ireland (IE), where the best alignment was obtained with a significantly higher NDCG threshold than in other countries. Consequently, its CLS score is also among the highest. This deviation can be explained by the shared official language (English) between Ireland and the UK, the target model. Since both vector spaces are already linguistically and semantically close, fewer anchor points are needed to achieve a high-quality alignment, reducing the need for a lower threshold.

These results further highlight the importance of carefully selecting the NDCG threshold for alignment. While a lower threshold generally improves alignment, language similarities between the source and target may allow for effective alignments even with fewer anchor points.

5. Assessing skill bundles across Europe using `VEUCTOR`

In the previous sections, we constructed various embeddings and identified the best- and worst-aligned embeddings after training, aligning, and evaluating all models. In this section, we show how alignment quality affects occupational representations and the relationships between occupations and skills.

5.1. Occupation representation

We first define how occupations are represented in the embedding space to analyze differences between the best- and worst-performing alignment models for a specific occupational task. Rather than relying solely on the occupation term itself, we construct a more informative representation by leveraging the associated skill vectors. In essence, different embedding models are likely to generate different skill bundles, which are then likely to yield different results for the analysis.

For each occupation, we compute a centroid vector by averaging the embeddings of all the skills related to that occupation. Given an occupation o with a set of associated skills $S_o = \{s_1, s_2, \ldots, s_n\}$, where each skill s_i has an embedding $v(s_i)$, the centroid of the

occupation is defined as:

$$v_o = \frac{1}{|S_o|} \sum_{s \in S_o} v(s) \tag{5}$$

In other words, v_o defines a vector representation of the skill bundle of occupation o. The vector representation is very useful, as it allows one to compute measures of differences and similarities.

5.2. Measuring skill similarities

To evaluate the potential bias introduced by model selection, we developed a simple indicator. Specifically, for each detailed occupation at the ISCO 4-digit level, we calculated a similarity measure for the associated skill bundle using the Jaccard distance between each identified skill and all other skills. To make the construction of the skill similarity indicator explicit, let o denote an occupation and let $S_o = \{s_1, \ldots, s_{n_o}\}$ be the set of skills associated with o. For any pair of occupations (o_i, o_j), the Jaccard similarity is defined as

$$J(o_i, o_j) = \frac{|S_{o_i} \cap S_{o_j}|}{|S_{o_i} \cup S_{o_j}|}. \tag{6}$$

Operationally, this formulation reduces to evaluating whether two skills co-occur within the same occupational skill bundle.

Given the high level of occupational specificity, we expect the most effective embeddings to generate, within occupations, more internally consistent (i.e., similar) skill bundles compared to the least effective embeddings. Consequently, the magnitude of the difference in occupation skill similarity measures between the best and worst embeddings indicates the extent of bias introduced by model selection.

As emphasized in the previous paragraphs, developing embeddings in a multilingual setting poses significant challenges, including selecting the optimal embeddings and aligning them for comparability. In principle, our results are affected by two elements: the optimization of embeddings and the optimization of the alignment procedure. In order to separate the two effects we start by performing an *intra*-country analysis, i.e., we compare *unaligned* models, focusing on differences between the best- and worst-performing models within each country. In this case, the analysis can be performed within countries, but the results cannot be compared between countries. Fig. 6 displays the results. The left panel plots the cumulative distribution of the skill similarity measure by occupation of the best (blue line) and worst (red line) embeddings for a selection of countries (IT, DE, UK, PL, ES). In both cases, the best embedding delivers a more similar set of skills by occupation. Beyond the eyeball metric offered by the two figures, we performed a formal test for the two distributions. The Goldman Kaplan test rejects the equality of the two CDFs at 1%. The region of rejection of the family-wise error rate is highlighted by the thick black line underlying the graph.[14] The right panel displays box plots of the values of skill similarity by major occupation groups (1 Digit ESCO.). Similarly, there is a notable difference in the distribution of skill similarity by occupation between the two models, with the best embeddings yielding higher overall skill similarities across occupations.[15]

Therefore, our analysis has shown that the choice of embeddings can lead to significantly different skill bundles, ultimately influencing the outcomes of the analysis. Up to this point, the embeddings generated for different countries have not been aligned, restricting comparisons to within-country skill bundles. Next, we introduce the *alignment* procedure to enable cross-country comparisons of the results. Since both embeddings and alignment introduce degrees of freedom, we systematically explored all possible combinations by generating different skill bundles under the best and worst embedding conditions, as well as the best and worst alignment parameters.

This analysis yields two key findings. First, the alignment model produces highly robust results across different parameter specifications. Statistical tests fail to reject, at any significance level, the null hypothesis that the distribution of skill bundles is identical across different alignment configurations. This indicates that parameter selection for the alignment model does not introduce meaningful bias in the results.

Second, in contrast, the choice of the embedding model proves to be of critical importance. Fig. 7 presents a comparative analysis of the distribution of skill bundles across countries, contrasting the best and worst embeddings. The corresponding boxplot illustrates these differences within ISCO 1-digit occupational groups. Consistent with the intra-country results, cross-country comparisons further confirm that the best embeddings yield more similar skill bundles than the worst embeddings.

Fig. 8 extends the cross-country analysis by comparing distributions across countries. The results demonstrate that the best embeddings consistently generate skill bundles that exhibit higher similarity across countries, reinforcing the robustness of this finding.

We conclude this section by providing some refinements to our analysis. We begin by considering not only the mean similarity across skill bundles but also its variance (Fig. 9 panel a). As expected, the comparison of the two distributions shows that the variance of similarity measures is higher for the worst model than for the best ones. In other words, the best models are characterised by more compact skill bundles, with both higher mean similarity and lower variance.

[14] Instead of testing a single global null hypothesis (that the two CDFs are identical), the Goldman Kaplan methodology tests a continuum of individual null hypotheses of CDF equality at each point. This allows us to obtain the ranges of x (if it exists) where $F(x) = G(x)$ is rejected at the specified error level. As the test is conducted for all points of the CDF, the methodology is exposed to the "multiple testing problem" that generates the "familywise error" where at least one true H_0 is rejected. Our approach achieves "strong control" of FWER, at a 5% level, meaning that there are zero false positives 95% of the time.

[15] The high variability of results for group 6 (Agricultural, Forestry and Fishery workers) is due to the low number of observations in OJAs for these occupations.

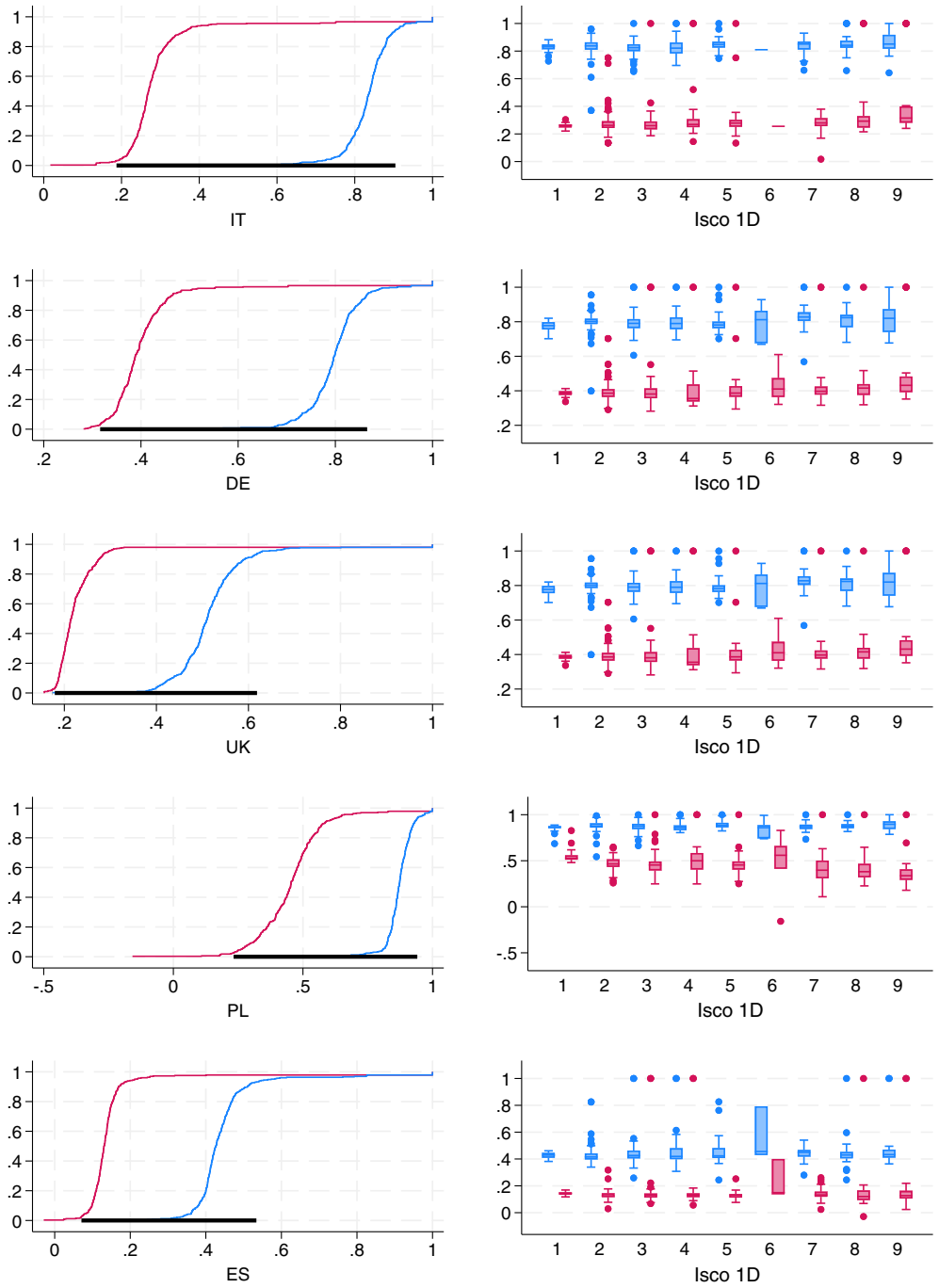
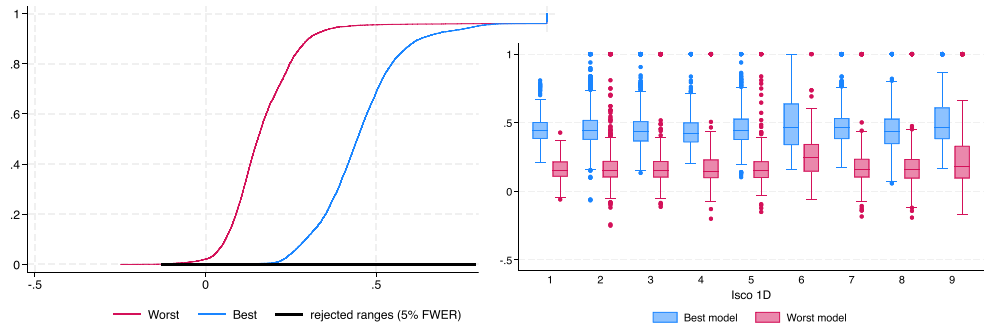


Fig. 6. Skill bundles comparison intra-country: best and worst embeddings. Non aligned models.
Best models: blue line/ boxplot. Worst model: red line/boxplot. Left panel: comparison of cumulative distributions of mean similarities of skill bundles generated with best and worst embeddings. The underlying black line denotes the region of rejection of the null hypothesis of equal distribution. Right panel: box plots of the distribution of mean similarity by occupation (1 digit ESCO)

Finally, we have split the skill set into three categories - hard, soft, and digital skills - and performed the analysis separately for each group.[16] Fig. 9, panels b, c, d show that the results obtained for the overall skill set are confirmed within each category: best models deliver more similar and statistically distinct skill bundles compared to the worst models.[17]

[16] The categories are constructed using the ESCO classification.

[17] Note that in the digital skill set, there is a concentration at the end of the cumulative distribution. This is expected, as the set of digital skills is limited, and several OJAs contain few general digital skills, such as Excel, where there is no difference between embedding models.



(a) Distribution comparison of best and worst models (b) Distribution of mean skill similarity over 1Digit ISCO

Fig. 7. Skill bundles comparison inter-country: best and worst embeddings. Aligned models.
Best models: blue line/ boxplot. Worst model: red line/boxplot. Left panel: comparison of cumulative distributions of mean similarities of skill bundles generated with best and worst embeddings. The underlying black line denotes the region of rejection of the null hypothesis of equal distribution. Right panel: box plots of the distribution of mean similarity by occupation (1 digit ESCO)

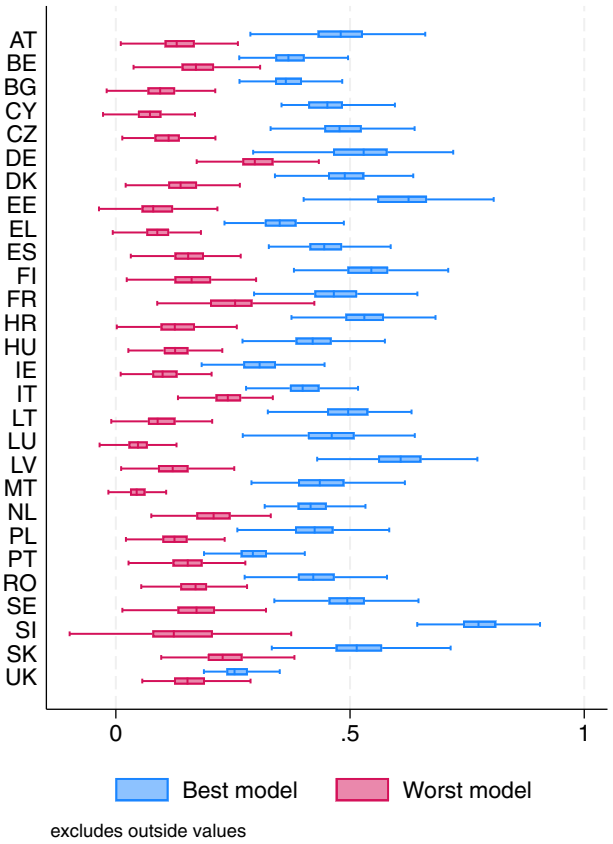


Fig. 8. Distribution of mean skill similarity by country. Inter country comparison. Aligned models.

Overall, our results show that the selection of the embedding model generates a substantial and statistically significant variation in the distribution of skill bundles. On the contrary, the fine-tuning of alignment configurations does not have any statistically relevant effect.

5.3. Extrinsic evaluation

In the previous section, we showed that different embedding models lead to sizeable differences in the resulting skill-bundle representations. In this section, we go one step further by assessing whether these alternative embeddings also generate meaningful differences in downstream applications. In fact, although our embedding selection task is intrinsically task-independent - as it is

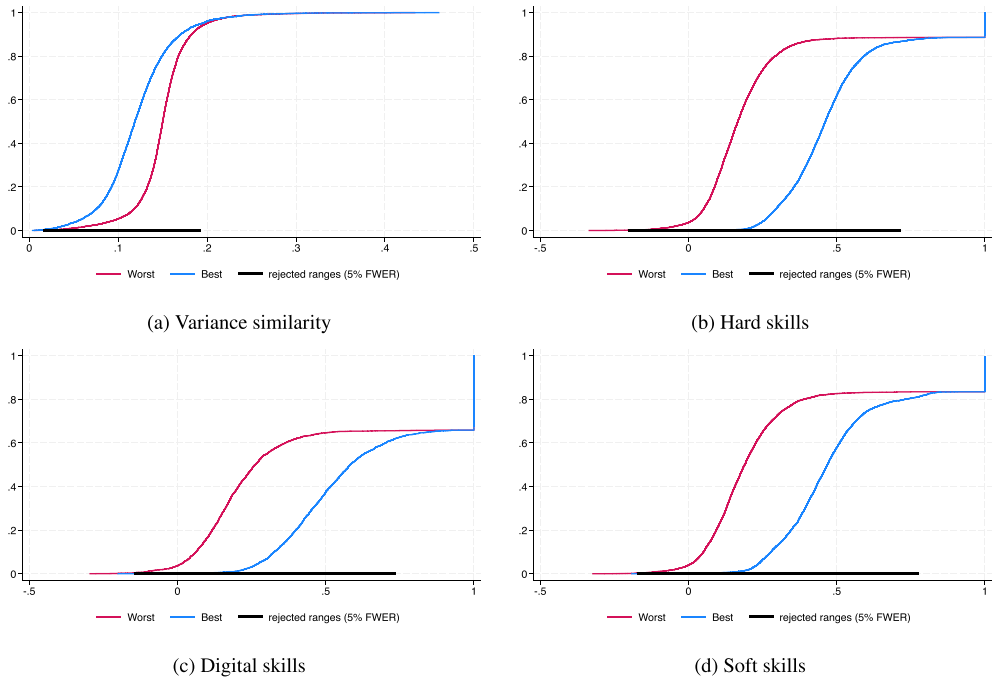


Fig. 9. Skill bundles comparison inter-country: best and worst embeddings. Aligned models.
Best model: blue line. Worst model: red line. Comparison of cumulative distributions of variance (panel a) and mean similarities (panels b, c, d) of skill bundles generated with best and worst embeddings. The underlying black line denotes the region of rejection of the null hypothesis of equal distribution.

designed to fit the entire structure of ESCO rather than being tailored to a specific downstream application - we nevertheless obtain strong confirmation of the validity of our approach through extrinsic evaluation.

We consider two different extrinsic tasks: the OJAs classification into the 4-digit ESCO occupations (436 classes) and the prediction of standard labor market variables. The goal of these experiments is to verify whether the embeddings identified by our methodology as best and worst performing models lead to systematically different outcomes when employed in a real-world task. In other words, this exercise serves as an extrinsic validation of the proposed embedding selection framework.

As the classification evaluation necessitates a benchmark, we have selected 4 countries (IT, DE, FI, UK) for which a gold benchmark was available. The benchmark was created through a manual annotation process carried out by labor market experts.[18] These experts were asked to label a sample of job advertisements with the most appropriate occupation according to their domain knowledge and national labor market context.[19]

For each country, we have replicated the methodology described in this paper. We have trained a pool of FastText models using the same hyperparameter grid described in Section 4.2, restricting the training corpus to job advertisement titles only. This choice is motivated by two sets of considerations. First, the construction of the gold benchmark relies on OJA titles; therefore, to ensure a coherent comparison, model training must be conducted on the same set of information. Second, the literature provides evidence (see, e.g., [8]) that training embeddings directly on job-related textual content yields more reliable semantic representations for labor market applications, while avoiding potential information leakage from structured metadata.

Using the methodology described in Section 3, the best- and worst-performing models were identified for each country. These models were then evaluated in an occupational classification task. Classification was carried out by computing the cosine similarity between the title embedding and the embeddings associated with occupational labels, assigning the occupation corresponding to the highest similarity (top-1 prediction). This deliberately simple classification strategy was adopted to isolate the effect of embedding quality rather than introducing additional modelling complexity.

Table 5 reports the classification accuracy obtained using the best and worst performing embeddings for the countries considered. The results show a clear and consistent pattern: in all cases, the embedding selected as best by our methodology substantially outperforms the corresponding worst model. In some cases, such as Finland, the performance gap is particularly pronounced, highlighting the practical impact of embedding selection on downstream tasks.

[18] Note that these countries not only cover a large share of the online job advertisements available in the Web Intelligence Hub, but also represent substantial linguistic diversity.

[19] The details of the methodology are available from Eurostat at `https://cros.ec.europa.eu/wih`.

Table 5
Extrinsic evaluation results. For each country, the table reports the hyperparameters of the best and worst embedding models selected, together with their classification accuracy and the accuracy difference (best minus worst).

Country	Model	v_{size}	τ	**A**	hs	α	Accuracy	Δ Acc.
DE	Best	300	50	SG	✓	0.10	0.559	+0.158
	Worst	50	10	CBOW	✗	0.01	0.401	
FI	Best	300	100	SG	✓	0.05	0.627	+0.479
	Worst	50	10	SG	✗	0.01	0.148	
IT	Best	300	100	SG	✗	0.10	0.534	+0.128
	Worst	50	10	SG	✗	0.01	0.406	
UK	Best	300	10	CBOW	✗	0.01	0.673	+0.026
	Worst	300	100	CBOW	✓	0.10	0.647	

Table 6
RMSE and R2 comparison.

	Wage	Education	Experience	Contract
RMSE (↓)				
FastText Best	**1.4361**	**0.6951**	**0.608**	**0.4338**
FastText Worst	1.4692	0.7023	0.6294	0.4531
bge-large-en-v1.5	1.6280	0.7565	0.7275	0.4775
bilingual-embedding-base	1.6285	0.7567	0.7279	0.4777
gte-large	1.6287	0.7566	0.7272	0.4773
multilingual-e5-large-instruct	1.6286	0.7566	0.7276	0.4775
R2 (↑)				
FastText Best	**0.2162**	**0.1825**	**0.1711**	**0.1651**
FastText Worst	0.1797	0.1775	0.1675	0.0889
bge-large-en-v1.5	0.1429	0.1340	0.1319	0.0720
bilingual-embedding-base	0.1424	0.1335	0.1310	0.0716
gte-large	0.1422	0.1338	0.1326	0.0729
multilingual-e5-large-instruct	0.1424	0.1338	0.1317	0.0723

The second downstream task that we considered is the prediction of labor market information contained in OJAs. Specifically, we have used the information content of the skill bundles for hard, digital, and soft skills[20] to predict wages, education, experience, and contract type. More precisely, we exploit the information available in the Web Intelligence Hub on wages, education levels, years of experience, and types of working contracts associated with each OJA. Since the representation of the skill bundles that we provide in the paper is constructed at the occupation level, we compute the occupation level average of each variable and regress it on the coefficient of variation of the skill bundles derived for hard, digital, and soft skills.

The rationale for this exercise is to assess whether the information content embedded in the skill bundles generated by the Best and Worst models helps in predicting wages, education levels, experience, and contract types. Table 6 compares the Root Mean Squared Error (RMSE), and the R2 obtained from the two models, and shows that embeddings generated by Best models exhibit both higher explanatory power and higher predictive accuracy (higher values of R2 and lower values of the RMSE are reported in bold).[21,22]

These findings support the validity and practical relevance of the proposed methodology. The best models tend to outperform alternative ones in standard downstream applications, such as occupation classification or the prediction of standard economic variables. This reinforces the main message of the paper: the choice of word embeddings is not neutral, and systematic, taxonomy-driven selection can lead to more reliable and reproducible labor market intelligence.

5.4. Reproducibility

Code and data availability. The codes, trained embedding models, and their aligned versions are made publicly available for research purposes. We provide pre-trained FastText embeddings for multiple countries, along with their aligned counterparts. A GitLab repository has been created to facilitate reproducibility and further analysis. This repository contains the scripts and supplementary data used for evaluating both the original FastText models and their aligned versions.

[20] Essentially, the information contained in Fig. 9 of the paper.

[21] The table reports the analysis considering occupation fixed effects. We have conducted the analysis both without and with country fixed effects, obtaining analogous results.

[22] For the sake of completeness, Table 6 compares FastText Models with LLMs used in Table 3. Overall, embeddings generated by LLMs exhibit both lower explanatory power and lower predictive accuracy relative to the best-performing FastText models.

```python
import fasttext
model = fasttext.load_model('path/to/fasttext/model_best_uk')

from gensim.models.keyedvectors import Word2VecKeyedVectors
model = Word2VecKeyedVectors.load_word2vec_format('path/to/aligned/model_best_uk')
```

Listing 1. Example of loading the best model in a Python environment for UK.

In the Listing 1, an example of loading both a FastText model and an aligned model is provided. The official fasttext[23] library is used to load the model. For the aligned model, the widely used gensim[24] library is employed to load and utilize the model. Examples are provided using Python to align with the code used in our experiments; however, the choice of programming language is flexible. The FastText models are in `.bin` format, which can be loaded using various languages (e.g., Python, MATLAB, R, etc.), and the aligned models are provided in `.txt` format, making them even easier to handle.

All code used in this study was written in Python 3.12. The complete implementation is available in a Github repository (https://github.com/Crisp-Unimib/veuctor), which provides the scripts and additional data needed to reproduce our experiments.

Within the repository, the data folder contains supplementary resources, including the ESCO taxonomy for occupations and skills. Two Python scripts are included: `HSS_eval`, which implements the evaluation of the generated models using the HSS, and `alignment_eval`, which evaluates the aligned models. Additionally, a demo script is provided to illustrate the use of both models and the evaluation scripts. The repository also includes a README.md file with detailed instructions on usage and setup.

Hardware specifications and processing time. The computational experiments conducted in this study required significant processing power due to the large-scale training and alignment of embedding models across multiple countries. All computations were performed on an AWS cloud infrastructure to ensure efficiency and reproducibility.

The training of both the embedding models and their alignments was conducted using three machines,[25] each equipped with 32 CPUs and 64 GB of memory, along with one high-performance machine featuring 64 CPUs and 128 GB of memory.[26] The latter was specifically used for training models on larger corpora, such as those from the UK, France, and Germany. Training was parallelized both within each machine - distributing processes across available CPUs - and across multiple machines, with different countries assigned to different systems to optimize efficiency. All machines are equipped with an AMD EPYC 7R13 processor.

From Table 2, we observe that training times varied significantly across countries due to differences in corpus size and computational complexity. Countries with larger OJAs corpora, such as the UK, France, and Germany, required substantially more time for model convergence. Additionally, the total training time for all models combined exceeded 140 days, highlighting the scale of the computational effort involved.

The alignment process was considerably faster than training, as it involved matrix transformations rather than learning word representations from scratch. The generation and evaluation of the aligned model pool were completed in approximately 2 days and 10 hours.

6. Concluding remarks

This paper shows that the choice of word embeddings has a substantial impact on the information extracted from both structured and unstructured text. In particular, embedding models introduce systematic biases that directly shape downstream analyses, making their selection a critical methodological decision rather than a neutral preprocessing step.

We have developed a methodology to evaluate and compare different word embeddings, identifying the most effective one for a given task. Additionally, we introduced a framework for aligning word embeddings across multiple languages, enhancing their applicability in multilingual contexts.

Applying our methodology to a comprehensive dataset of Online Job Advertisements in Europe, we showed that different word embeddings yield distinct informational outputs, leading to substantially different skill bundle categorizations. As a key contribution to the research community, we provide `VEUCTOR`, a tool that not only implements our methodology but also offers access to precomputed word embeddings. This resource enables researchers to conduct labor market analyses with an optimized information extraction approach.

Although our study focuses on the European labor market, our findings have broader implications, extending to any domain that relies on embeddings to extract insights from textual data. This is particularly relevant in the social sciences, where the increasing availability of large-scale unstructured text has driven a proliferation of analytical tools and research methodologies. We, therefore, advocate for the adoption of our approach in various fields, facilitating more accurate and context-aware analyses.

Limitations. While our work is applied to the EU labor market, we build a general methodological approach that can be extended to several other domains. In order to do so, some caveats need to be stressed.

[23] https://fasttext.cc/docs/en/python-module.html.
[24] https://radimrehurek.com/gensim/.
[25] AWS 3x c6a.8xlarge.
[26] AWS c6a.16xlarge.

First, the effectiveness and clarity of the proposed model selection framework depend on the availability of a clear, robust, and semantically grounded benchmark. In our case, within the context of skill intelligence, we can rely on the ESCO taxonomy, which provides a natural and well-established reference for evaluating embedding quality. However, in other application domains, such an explicit and widely accepted benchmark may be unavailable or only partially defined, potentially limiting the direct applicability of the proposed evaluation strategy.

Second, our approach highlights an inherent tension between intrinsic and extrinsic evaluation. From a methodological perspective, intrinsic evaluation against a reliable benchmark offers a controlled and principled means of comparing embedding models and identifying optimal configurations. Nonetheless, embedding quality, as measured intrinsically, does not automatically translate into superior performance across every downstream task. While our experiments show consistent advantages in several extrinsic applications, the relationship between intrinsic optimality and task-specific performance remains context-dependent and may vary across domains.

Third, our analysis relies on a temporally fixed snapshot of Online Job Advertisement data, and therefore reflects labor market structures at a specific point in time. As a consequence, the results should not be interpreted as capturing temporal dynamics or structural changes in skill demand. This limitation is primarily driven by data availability constraints, as access to updated OJA-WIH data is governed by Eurostat data release and access policies. Importantly, this is not a limitation of the proposed methodology itself: once new data become available, the embedding models can be retrained and re-aligned using the same pipeline. The framework is thus inherently updatable and suitable for periodic extensions as new data releases occur.

CRediT authorship contribution statement

Emilio Colombo: Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Simone D'Amico:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. **Fabio Mercorio:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Mario Mezzanzanica:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Algorithm 1 focuses on generating and evaluating word embedding models for various countries based on online job advertisements (OJAs). The primary objective is to create embedding models that accurately reflect occupational semantics within each country's labor market. This process is divided into two major steps: the training of embeddings and their subsequent evaluation.

The first step involves the preprocessing of OJA corpora specific to each country to ensure consistency and the removal of noise. This includes standard text cleaning procedures such as normalization, stop-word removal, and tokenization. Once the data is preprocessed, multiple `FastText` models are trained using varying hyperparameters, including embedding size, number of epochs, learning rate, the choice between Skip-Gram and CBOW algorithms, and the use of hierarchical softmax. This extensive grid search over hyperparameters ensures that the models can capture the nuances specific to each country's labor market data.

The second step focuses on evaluating the trained embeddings. This is achieved by comparing the semantic relationships captured by the embeddings against the ESCO taxonomy, a structured representation of European occupations and skills. Two key metrics are used in this evaluation: cosine similarity, which measures the closeness of vector representations, and the Hierarchical Skill Similarity (HSS), which assesses taxonomic proximity. By computing the Spearman rank correlation between these two measures, the algorithm identifies how well the embedding model preserves the occupational relationships defined in the ESCO taxonomy.

An essential feature of this evaluation process is its ability to distinguish between the best and worst-performing models for each country. The best model is the one that achieves the highest Spearman correlation, indicating a strong alignment between the learned embeddings and the taxonomic structure. Conversely, the worst model has the lowest correlation, often revealing gaps between the real labor market dynamics and the theoretical taxonomy. This step not only highlights the variability in model performance across countries but also underscores the challenges in creating universally high-quality embeddings.

Algorithm 2 builds upon the embeddings generated in Algorithm 1 by addressing the challenge of cross-country comparability. While Algorithm 1 ensures that embeddings accurately reflect national labor market structures, these models are inherently non-comparable across countries due to differences in languages, training data, and contextual nuances. Algorithm 2 resolves this by aligning country-specific embeddings into a shared vector space, enabling meaningful cross-lingual and cross-country analyses.

The alignment process employs the `SeNSe` technique, which maps the source embeddings from each country onto a common target space - in this case, the UK model - chosen as the reference point. This alignment uses a set of anchor points, typically occupations that exist in both the source and target spaces, to perform a geometric transformation that preserves semantic relationships. The transformation is guided by hyperparameters, including the Normalized Discounted Cumulative Gain (NDCG) threshold, which controls the selection of anchor points.

Once the alignment is complete, the quality of the aligned embeddings is evaluated using the Cross-Lingual Semantic Fitting Score (CLS score). This score measures the degree to which the aligned vectors match the target model by calculating the cosine

Algorithm 1 Embedding pool training and evaluation (Step 1 & step 2).

Require: C ▷ Set of countries analyzed
Require: Θ ▷ Set of hyperparameter configurations for embedding models
Require: $\{\mathcal{D}_c\}_{c \in C}$ ▷ Corpus of OJAs for each country
Require: $\mathcal{O}$ ▷ Set of occupations in the ESCO taxonomy
Ensure: $\theta_C^{\top}$ ▷ Set of best parameter combinations for each country
Ensure: $\theta_C^{\perp}$ ▷ Set of worst parameter combinations for each country

$\theta_C^{\top} \leftarrow \emptyset$
$\theta_C^{\perp} \leftarrow \emptyset$
for $c \in C$ **do**
  for $\theta \in \Theta$ **do**
    $\mathcal{M}_{c,\theta} \leftarrow train(\mathcal{D}_c, \theta)$ ▷ Step 1. Train the model on the corpus of country c
    $\mathbf{S}_{cos} \leftarrow \emptyset$ ▷ Step 2. Stores cosine similarity scores
    $\mathbf{S}_{HSS} \leftarrow \emptyset$ ▷ Stores HSS values scores
    for each $(o_i, o_j) \in \mathcal{O} \times \mathcal{O}$ where $i > j$ **do**
      $v_{o_i} \leftarrow \mathcal{M}_{c,\theta}(o_i)$ ▷ Encoder occupations o_i and o_j
      $v_{o_j} \leftarrow \mathcal{M}_{c,\theta}(o_j)$
      $\mathbf{S}_{cos} \leftarrow \mathbf{S}_{cos} \cup cosine(v_{o_i}, v_{o_j})$
      $\mathbf{S}_{HSS} \leftarrow \mathbf{S}_{HSS} \cup HSS(o_i, o_j)$
    end for
    $\rho_{c,\theta} \leftarrow \text{spearman}(\mathbf{S}_{cos}, \mathbf{S}_{HSS})$
  end for
  $\theta_c^{\top} \leftarrow \arg\max_{\theta \in \Theta} \rho_{c,\theta}$ ▷ Best parameter combination for country c
  $\theta_c^{\perp} \leftarrow \arg\min_{\theta \in \Theta} \rho_{c,\theta}$ ▷ Worst parameter combination for country c
  $\theta_C^{\top} \leftarrow \theta_C^{\top} \cup \theta_c^{\top}$
  $\theta_C^{\perp} \leftarrow \theta_C^{\perp} \cup \theta_c^{\perp}$
end for

Algorithm 2 Alignment embedding pool training and evaluation (Step 3 & step 4).

Require: C ▷ Set of countries analyzed
Require: Θ ▷ Set of hyperparameter configurations for embedding models
Require: $\{\mathcal{M}_c\}_{c \in C}$ ▷ Source model for each country
Require: $\mathcal{M}_{target}$ ▷ Target model for the alignemnt
Require: $\mathcal{O}$ ▷ Set of occupations in the ESCO taxonomy
Ensure: $\theta_C^{\top}$ ▷ Set of best parameter combinations for each country
Ensure: $\theta_C^{\perp}$ ▷ Set of worst parameter combinations for each country

$\theta_C^{\top} \leftarrow \emptyset$
$\theta_C^{\perp} \leftarrow \emptyset$
for $c \in C$ **do**
  for $\theta \in \Theta$ **do**
    $\mathcal{A}_{c,\theta} \leftarrow SeNSe(\mathcal{M}_c, \mathcal{M}_{target}, \theta)$ ▷ Step 3. Align the source model of c to the target model
    $\mathbf{S}_{cos} \leftarrow \emptyset$ ▷ Step 4. Stores cosine similarity scores
    for each $o \in \mathcal{O}$ **do**
      $v_{\mathcal{A}_{c,\theta}} \leftarrow \mathcal{A}_{c,\theta}(o)$ ▷ Encoder occupation o
      $v_{\mathcal{M}_{target}} \leftarrow \mathcal{M}_{target}(o)$
      $\mathbf{S}_{cos} \leftarrow \mathbf{S}_{cos} \cup cosine(v_{\mathcal{A}_{c,\theta}}, v_{\mathcal{M}_{target}})$
    end for
    $CLS\_score_{\mathcal{A}_{c,\theta}} \leftarrow \text{mean}(\mathbf{S}_{cos})$
  end for
  $\theta_c^{\top} \leftarrow \arg\max_{\theta \in \Theta} CLS\_score_{\mathcal{A}_{c,\theta}}$ ▷ Best parameter combination for country c
  $\theta_c^{\perp} \leftarrow \arg\min_{\theta \in \Theta} CLS\_score_{\mathcal{A}_{c,\theta}}$ ▷ Worst parameter combination for country c
  $\theta_C^{\top} \leftarrow \theta_C^{\top} \cup \theta_c^{\top}$
  $\theta_C^{\perp} \leftarrow \theta_C^{\perp} \cup \theta_c^{\perp}$
end for

similarity between corresponding occupation vectors in the source and target spaces. A high CLS score indicates that the alignment has successfully preserved the semantic structure across languages and contexts.

Algorithm 2 also identifies the best and worst-aligned models based on the CLS score. Notably, the quality of alignment varies depending on linguistic proximity to the target model. For example, countries sharing the same language as the target (e.g., Ireland and the UK) naturally achieve better alignment with fewer anchor points. Conversely, more linguistically distant countries require a denser set of anchors to achieve comparable alignment quality.

Both algorithms underscore the importance of careful model selection and alignment in labor market analyses. Algorithm 1 highlights the importance of high-quality, country-specific embeddings that accurately reflect local occupational semantics. In contrast, Algorithm 2 highlights the complexities involved in aligning these embeddings across countries, especially in multilingual contexts, and the critical role of alignment quality in enabling robust cross-country comparisons.

Data availability

150GB + data were published through Zenodo links (reported in PDF) and made visible after paper publications.

References

[1] M. Artetxe, G. Labaka, E. Agirre, A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, arXiv preprint arXiv:1805.06297, 2018.
[2] J. Azar, E. Huet-Vaughn, I. Marinescu, B. Taska, T. von Wachter, Minimum wage employment effects and labor market concentration, Rev. Econ. Stud. (2023) rdad091, https://doi.org/10.1093/restud/rdad091
[3] J. Azar, I. Marinescu, M. Steinbaum, B. Taska, Concentration in US labor markets: evidence from online vacancy data, Labour Econ. 66 (2020) 101886, https://doi.org/10.1016/j.labeco.2020.101886
[4] I.M. Azpiazu, M.S. Pera, Hierarchical mapping for crosslingual word embedding alignment, Trans. Assoc. Comput. Linguist. 8 (2020) 361–376.
[5] A. Bhola, K. Halder, A. Prasad, M.-Y. Kan, Retrieving skills from job descriptions: a language model based extreme multi-label classification framework, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 5832–5842.
[6] L. Bjerkander, A. Glas, Talking in a language that everyone can understand? Clarity of speeches by the ECB executive board, J. Int. Money Finance 149 (2024) 103200, https://doi.org/10.1016/j.jimonfin.2024.103200
[7] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146.
[8] R. Boselli, M. Cesarini, F. Mercorio, M. Mezzanzanica, Classifying online job advertisements through machine learning, Futur. Gener. Comput. Syst. 86 (2018) 319–328.
[9] J.C. Braxton, B. Taska, Technological change and the consequences of job loss, Am. Econ. Rev. 113 (2023) 279–316, https://doi.org/10.1257/aer.20210182, https://www.aeaweb.org/articles?id=10.1257/aer.20210182.
[10] B. Clavié, G. Soulié, Large language models as batteries-included zero-shot esco skills matchers, arXiv preprint arXiv:2307.03539, 2023.
[11] E. Colombo, A. Marcato, Skill demand and labour market concentration: evidence from Italian vacancies, Int. J. Manpower 44 (2023) 156–198, https://doi.org/10.1108/IJM-04-2023-0181
[12] E. Colombo, F. Mercorio, M. Mezzanzanica, AI meets labor market: exploring the link between Automation and skills, Inf. Econ. Policy 47 (2019), https://doi.org/10.1016/j.infoecopol.2019.05.003, http://www.sciencedirect.com/science/article/pii/S0167624518301318.
[13] E. Colombo, F. Mercorio, M. Mezzanzanica, A. Serino, Towards the terminator economy: assessing job exposure to AI through llms, in: [object Object] (Ed.), Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, International Joint Conferences on Artificial Intelligence Organization, 2025, pp. 9591–9600, https://doi.org/10.24963/ijcai.2025/1066. AI and Social Good.
[14] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, H. Jégou, Word translation without parallel data, arXiv preprint arXiv:1710.04087, 2017.
[15] A. De Santo, L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, Skillens: recognising and mapping novel skills from millions of job ads across Europe using language models, in: The 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026) (To Appear), 2026.
[16] J.-J. Decorte, J. Van Hautte, T. Demeester, C. Develder, Jobbert: Understanding job titles through skills, arXiv preprint arXiv:2109.09605, 2021.
[17] J.-J. Decorte, S. Verlinden, J. Van Hautte, J. Deleu, C. Develder, T. Demeester, Extreme multi-label skill extraction training using large language models, arXiv preprint arXiv:2307.10778, 2023.
[18] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, GrapHLMI: a data driven system for exploring labor market information through graph databases, Multim. Tools Appl. 81 (2022) 3061–3090, https://doi.org/10.1007/S11042-020-09115-X
[19] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, Weta: automatic taxonomy alignment via word embeddings, Comput. Ind. 138 (2022) 103626.
[20] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, A. Seveso, Neo: a tool for taxonomy enrichment with new emerging occupations, in: International Semantic Web Conference, Springer, 2020, pp. 568–584.
[21] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, A. Seveso, Skills2job: a recommender system that encodes job offer embeddings on graph databases, Appl. Soft Comput. 101 (2021) 107049.
[22] A. Goldfarb, B. Taska, F. Teodoridis, Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings, Res. Policy 52 (2023) 104653, https://doi.org/10.1016/j.respol.2022.104653
[23] R. Gu, L. Zhong, Effects of stay-at-home orders on skill requirements in vacancy postings, Labour Econ. 82 (2023) 102342, https://doi.org/10.1016/j.labeco.2023.102342
[24] S. Hansen, M. McMahon, A. Prat, Transparency and deliberation within the FOMC: a computational linguistics approach*, Q. J. Econ. 133 (2017) 801–870, https://doi.org/10.1093/qje/qjx045
[25] B. Hershbein, L.B. Kahn, Do recessions accelerate routine-biased technological change? Evidence from vacancy postings, Am. Econ. Rev. 108 (2018) 1737–1772, https://doi.org/10.1257/aer.20161570
[26] C. Leacock, M. Chodorow, G.A. Miller, Using corpus statistics and Wordnet relations for sense identification, Comput. Linguist. 24 (1998) 147–165.
[27] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards general text embeddings with multi-stage contrastive learning, arXiv preprint arXiv:2308.03281, 2023.
[28] A. Maedche, S. Staab, Ontology learning for the semantic web, IEEE Intell. Syst. 16 (2001) 72–79.
[29] L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, MEET-LM: a method for embeddings evaluation for taxonomic data in the labour market, Comput. Ind. 124 (2021) 103341, https://doi.org/10.1016/J.COMPIND.2020.103341
[30] L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, Taxoref: embeddings evaluation for ai-driven taxonomy refinement, in: ECML PKDD, 2021.
[31] L. Malandri, F. Mercorio, M. Mezzanzanica, F. Pallucchini, Sense: embedding alignment via semantic anchors selection, Int. J. Data Sci. Anal. (2024) 1–15.
[32] L. Malandri, F. Mercorio, A. Serino, Skillmo: normalized ESCO skill extraction through transformer models, in: J. Hong, S. Battiato, C. Esposito, J.W. Park, A. Przybylek (Eds.), Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, SAC 2025, Catania International Airport, Catania, Italy, 31 March 2025 - 4 April 2025, ACM, 2025, pp. 1969–1978, https://doi.org/10.1145/3672608.3707960
[33] M. Henning, R. Eriksson, P. Garefelt, H. Martin, Z. Elekes, Job relatedness, local skill coherence and economic performance: a job postings approach, Reg. Stud. Reg. Sci. 12 (2025) 95–122, https://doi.org/10.1080/21681376.2025.2459148

[34] M. Mezzanzanica, F. Mercorio, Big data enables labor market intelligence, In S. Sakr, A.Y. Zomaya (Eds.), Encyclopedia of Big Data Technologies, Springer, 2019, https://doi.org/10.1007/978-3-319-63962-8_276-1
[35] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013.
[36] T. Mikolov, Q.V. Le, I. Sutskever, Exploiting similarities among languages for machine translation, arXiv preprint arXiv:1309.4168, 2013.
[37] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, Mteb: Massive text embedding benchmark, arXiv preprint arXiv:2210.07316, 2022.
[38] M. Papoutsoglou, A. Ampatzoglou, N. Mittas, L. Angelis, Extracting knowledge from on-line sources for software engineering labor market: a mapping study, IEEE Access 7 (2019) 157595–157613.
[39] P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, J. Artif. Intell. Res. 11 (1999) 95–130.
[40] J. Rosenberger, L. Wolfrum, S. Weinzierl, M. Kraus, P. Zschech, Careerbert: matching resumes to ESCO jobs in a shared embedding space for generic job recommendations, Expert Syst. Appl. 275 (2025) 127043.
[41] S. Ruder, I. Vulić, A. Søgaard, A survey of cross-lingual word embedding models, J. Artif. Intell. Res. 65 (2019) 569–631.
[42] N. Seco, T. Veale, J. Hayes, An intrinsic information content metric for semantic similarity in Wordnet, in: Ecai, 2004, pp. 1089.
[43] N. Thakur, N. Reimers, J. Daxenberger, I. Gurevych, Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks, arXiv e-prints arXiv:2010.08240, 2020.
[44] M. Vinel, I. Ryazanov, D. Botov, I. Nikolaev, Experimental comparison of unsupervised approaches in the task of separating specializations within professions in job vacancies, in: Conference on Artificial Intelligence and Natural Language, Springer, 2019, pp. 99–112.
[45] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Improving text embeddings with large language models, arXiv preprint arXiv:2401.00368, 2023.
[46] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, arXiv:2309.07597, 2023.
[47] M. Zhang, R. Van Der Goot, B. Plank, Escoxlm-R: Multilingual taxonomy-driven pre-training for the job market domain, arXiv preprint arXiv:2305.12092, 2023.

Author biography

Emilio Colombo (M.Sc., Ph.D., University of Southampton) is Professor of Economics at Università Cattolica del Sacro Cuore. Previously, he held positions at the University of Milano-Bicocca. He is the author of numerous articles published in leading international journals and has taught and delivered more than 50 seminars at universities worldwide. His research interests include international economics, macroeconomics, applied economics, and labor economics. He has led several research projects focused on skill development and labor market analysis. Currently, he serves as the project leader of a major international initiative, funded by Eurostat and Cedefop, aimed at developing a European system for collecting and analyzing online job advertisements.

Simone D'Amico received his Master's degree in Data Science from the University of Milano-Bicocca, Italy, in 2021. He is currently pursuing a Ph.D. at the same university within the Big Data & Analytics for Business curriculum, while also working as a researcher at the University of Naples Federico II, Italy. His research interests include Natural Language Processing, with a focus on large language models, as well as data mining and information extraction techniques. He is also active in the field of Finance, applying multi-modal models and forecasting models to predict stock market trends.

Fabio Mercorio is a Full Professor of AI and Data Science at the University of Milan-Bicocca, and director of the Master in AI and Data Analytics at the same university. His research interests include artificial intelligence, eXplainable AI, and data science (vector-space models, LLMs) to support human decision-making. He has been involved in many national and international research projects as both PI and Senior Expert on putting AI and Big Data into practice. He has led multiple research projects on skill development and labor market analysis. He currently leads the research unit, funded by Eurostat and Cedefop, to develop a European system for analysing online job advertisements. He co-authored more than 100 papers.

Mario Mezzanzanica is a Full Professor of Information Systems at the University of Milan Bicocca and director of the Department of Statistics and Quantitative Methods. He is also the Scientific Director of the CRISP center and Vice-Rector for Advanced Training and Job Placement. His research interests include Big Data Analytics, Artificial Intelligence, Business Intelligence, and Knowledge Discovery. He has also been involved in several technical and scientific committees established by Public Institutions, aimed at studying new models and methodologies for designing, monitoring, and evaluating innovation projects with significant impacts on ICT-based public services. He has coordinated numerous national and international research projects and has published over 100 scientific works. He leads an international project funded by Eurostat and Cedefop to develop a the Web Intelligence Hub (WIH) for analysing online job ads, building on his experience in skill development and labor market research.