UNIVERSITÀ CATTOLICA del Sacro Cuore

Dipartimento di Economia Internazionale, delle Istituzioni e dello Sviluppo (DISEIS)

WORKING PAPER SERIES

WP 2601

February 2026

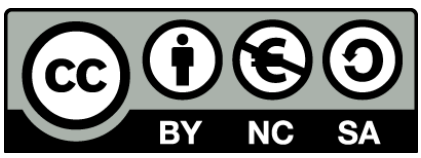
VEUCTOR : Training and Selecting Best Vector Space Models from Online Job Ads for European Countries

Emilio Colombo, Simone D'Amico, Fabio Mercorio, Mario Mezzanzanica

Prior to being published, DISEIS working papers are subject to an anonymous refereeing process by members of the editorial board or by members of the department.

DISEIS Working Papers often represent preliminary work which are circulated for discussion and comment purposes. Citation and use of such a paper should account for its provisional character. A revised version may be available directly from the author.

DISEIS WP 2601



DISEIS
Dipartimento di Economia Internazionale, delle Istituzioni e dello Sviluppo
Università Cattolica del Sacro Cuore
Via Necchi 5
20123
Milano

VEUCTOR: Training and Selecting Best Vector Space Models from Online Job Ads for European Countries*

Emilio Colombo[†]
Università Cattolica
del Sacro Cuore
CRISP

Simone D'Amico [‡]
CRISP and DISMEQ
University of Milano-Bicocca

Fabio Mercorio [§]
CRISP and DISMEQ
University of Milano-Bicocca

Mario Mezzanzanica [¶]
CRISP and DISMEQ
University of Milano-Bicocca

January 2026

*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/

[†]Email: emilio.colombo@unicatt.it.
[‡]Email: simone.damico@unimib.it
[§]Email: fabio.mercorio@unimib.it
[¶]Email: mario.mezzanzanica@unimib.it

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 3,000 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.

Keywords: word embedding, machine learning, labor market, NLP

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 labour-related data —including labour shortages and surpluses reports and skills online job ads analysis tools — from agencies such as Eurostat, Cedefop, Eurofound, and the European Labour 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 present 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. The first challenge has been addressed by O*NET 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—like 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 allow machines to reason on 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 the identi-

[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

fication of 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 3,000 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 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.

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 *dramatic* 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

[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/

amount of unstructured information contained in texts has led to a proliferation of tools, analyses, and research that process and analyze such information. 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 (Bjerkander and Glas, 2024; Hansen et al., 2017)). 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:

1. 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.

2. 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.

3. 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.

4. 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.

5. 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 the data for best/worst generated embeddings and their optimality score.

The next section describes in more detail the above-mentioned steps.

2 Preliminaries

2.1 Online Job Advertisements for Labor and Skill Intelligence.

In recent years, the intermediation by specialized web portals and services 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., (Turrell et al., 2018; Papoutsoglou et al., 2019; Javed et al., 2017; UK Commission for Employment and Skills, 2015; Vinel et al., 2019; Giabelli et al., 2021a,b)). 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. Online Job Advertisements (OJAs) possess these features and have become increasingly important for academic research and the development of innovative statistics in the last years. Academic research has exploited the granularity of OJAs to analyze labor market concentration (Azar et al., 2020, 2023; Hershbein and Kahn, 2018; Colombo and Marcato, 2023), but most importantly has leveraged the information contained in the text of the advertisement to analyze firms' skill requirements (Braxton and Taska, 2023; Gu and Zhong, 2023; Goldfarb et al., 2023; Modestino et al., 2020; Colombo et al., 2019; Martin Henning and Elekes, 2025). As stated in the introduction, information extraction from textual data relies crucially on Word embeddings; moreover, most of the research is conducted in the US, due to the lack of available methodological approaches able to deliver robust results in a multi-language 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 (**?**). The result of this effort is the Web Intel-

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

ligence 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 (Boselli et al., 2018a, 2017; Colombo et al., 2019; **?**; Colace et al., 2019). This study takes advantage of this initiative by using the knowledge base collected by the WIH to train and optimize the word embeddings.

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 and allow words to be manipulated mathematically, facilitating operations such as similarity measurement, clustering, and classification. Early techniques for word representation, known as frequency-based models, relied on counting word occurrences 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 (Harris, 1954). 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 major breakthrough in this area came with neural network-based embedding models, particularly Word2Vec (Mikolov et al., 2013a). This method employs a shallow, two-layer neural network to learn word representations through two alternative architectures: *Continuous Bag of Words* (CBOW) and *Skip-Gram* (SG). CBOW predicts a target word based on its surrounding context, making it computationally efficient and suitable for large datasets. Skip-Gram, instead, predicts context words given a target word, which allows it to better capture relationships between rare words by generating more training examples from a given corpus. These inno-

vations significantly improved the ability to model semantic similarity and paved the way for more advanced word embedding techniques.

Bojanowski et al. (2017) 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.

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* has emerged as a powerful alternative. These models, exemplified by transformer-based architectures such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018), generate word representations dynamically based on the surrounding context in a sentence. This allows them to differentiate between different meanings of the same word and better 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.

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 (Boselli et al., 2018b). Research has shown that models trained on job vacancies can effectively tackle various challenges in this field.

One notable application is job classification, where embeddings have been used to map job postings to standardized taxonomies. *JobBERT* by Decorte et al. (2021) is

a model fine-tuned on BERT with job-related texts to enhance classification performance. Similarly, Zhang et al. (2023) leverage a multilingual pre-training model to improve job classification across different languages, integrating domain knowledge from job taxonomies.

Another critical task is skill extraction from job descriptions. Bhola et al. (2020) introduces a deep learning approach to identify relevant skills associated with job postings, addressing the extreme multi-label nature of this problem. Beyond classification and extraction, machine learning models have been employed to generate synthetic job postings and skill datasets. Clavié and Soulié (2023) explores how large-scale language models can be used to match skills to job descriptions without requiring extensive labeled data. Additionally, Decorte et al. (2023) 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 improving labor market analytics, facilitating better 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 (Godfray, 2002; Giabelli et al., 2020, 2022; D'Amico et al., 2024). According to Maedche and Staab (2001), a taxonomy can be defined as:

[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 in various domains where they are used to structure knowledge in ontologies and controlled vocabularies. In Natural Language Processing (NLP) and labor market analysis, taxonomies such as ESCO[5] (European Skills,

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

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.

There are several measures 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 in (Leacock et al., 1998). 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 (Jiang and Conrath, 1997; Seco et al., 2004; Lin et al., 1998).

2.4 The Hierarchical Semantic Similarity at a glance

For our studies, we use the Hierarchical Semantic Similarity (HSS) developed by Giabelli et al. (2020) as a similarity measure. The measure has then been implemented as a tool to perform taxonomy refinement (Malandri et al., 2021). The HSS has been designed to measure how closely 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 $c \in \mathcal{C}$ which 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 concept c. Following (Resnik, 1999), one can supplement the taxonomy with a probability measure $p : \mathcal{C} \rightarrow [0, 1]$ such that for every concept $c \in \mathcal{C}$, $p(c)$ 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 $p(c_1) \leq p(c_2)$. To estimate the values of p, (Resnik, 1999) employs the frequency of concepts within a large

text corpus. However, we aim to infer the similarity values inherent to 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 the value of 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 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 granularity, thus respecting our definition of p.

Given two words w_1 and w_2, (Resnik, 1999) defines $c_1 \in s(w_1)$ and $c_2 \in s(w_2)$ 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., 2020) 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:

[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:

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

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

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

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 word w_1 and w_2 which have L as LCA. 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. (2019). 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. (2013b)

observed that word vectors trained on monolingual data exhibit comparable topological structures across different languages (Mikolov et al., 2013b). This observation laid the foundation for early alignment methods, which assumed that embedding spaces could be mapped through a simple linear transformation (Mikolov et al., 2013b). However, this assumption has 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, ranging from offline linear transformations, like Mikolov et al. (2013b), to online multilingual embedding adjustments as Chi et al. (2021). 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. (2018) 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 (2020), 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. (2017), 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 optimisation. Their method also introduces a novel validation metric and the CSLS similarity measure, achieving re-

sults comparable to supervised approaches, even for distant and low-resource language pairs. Among these approaches, SeNSe by Malandri et al. (2024) 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, it has been tested on four different languages (Italian, Spanish, Finnish, and German) aligned to English, achieving performance superior to state-of-the-art methods. This demonstrates its effectiveness in adapting multilingual vector spaces while preserving semantic consistency. Second, it has been applied to a real-world labor market scenario, where vector spaces were generated from different corpora analysis stepsa set of occupations was evaluated after transformation. This showcases its practical utility for cross-country occupational 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) train different models for each country, in which semantic 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 best and worst training parameter combinations by evaluating how well each model reflects the semantic structure of the ESCO taxonomy. 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 (**?**). 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 deter-

mined 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 sets of hyperparameters, 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 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 com-

pute 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 $\{\mathcal{M}_c\}c \in \mathcal{C}$ (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 of the same occupation in the target model and the aligned model. 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 ISCO08 standard. The second is the Skills pillar, which lists and describes competencies/skills which 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 on-line job market, a job advertisement is a document containing two main text fields: a title and a full description. The title shortly summarises the job position, while the full description field usually includes the position details and the relevant skills the employee must possess (see, e.g., Mezzanzanica and Mercorio (2019)). The OJAs here used 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 whole OJA dataset, which is today composed of 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 aims to leverage the richness of information extracted from online OJA portals. The goal of Natural Language Processing (NLP) dataflows is to use the raw job title and description 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, 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, for a total number of observations equal to 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 quality scope. 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 strip-

ping 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[6] 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 and retain 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 illustrate the dataset's heterogeneity and emphasize the importance of tailored preprocessing to ensure the extraction of meaningful patterns and improve the performance of subsequent word embedding models for capturing semantic relationships and domain-specific concepts effectively.

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, FastText models are trained on job advertisement data from 28 different countries to generate a diverse pool of embeddings for subsequent analysis. For each country, a separate corpus of job advertisements is processed, ensuring that linguistic and contextual differences are captured. The key component of this process involves hyperparameters optimization to assess how their variations impact

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

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

2*Country	2*Country Code	Number of OJAs			2*Languages	2*Occupations	2*Vocabulary Size	Text Lengths		
		2020	2021	Total			(# of words)	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					

the quality of the resulting word embeddings. A parameter grid search is employed to systematically explore different combinations of 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 strike 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 give the model 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 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, leading to faster training times and better performance on larger datasets.

Hierarchical softmax (hs): A parameter that 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 but may require more epochs to reach optimal performance. A higher learning rate allows for faster updates, which can speed up training but may lead to overshooting the optimal solution or instability in the learning process. The choice of learning rate significantly influences 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: $v_{size} \in \{50, 100, 300\} \times \tau \in \{10, 50, 100\} \times \mathbf{A} \in \{SG, CBOW\} \times hs \in \{0, 1\} \times \alpha \in \{0.01, 0.05, 0.1\}$.

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 there is no single parameter combination that consistently appears as the best across the different countries, as the performance depends both on the chosen parameters and the training corpus. For example, for the selected algorithm, Skip-gram (SG) seems to perform the best in the majority of cases (18 countries), while Continuous Bag of Words (CBOW) appears to be the 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 reflect the semantic relationships inherent in ESCO taxonomy fully.

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

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 select appropriate pre-trained models for comparison, we relied on the benchmark constructed by (Muennighoff et al., 2022), which provides a comprehensive evaluation of text embedding models. Their framework, the Massive Text Embedding Benchmark (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[7]. According to the leaderboard, we select 4 open-source pre-trained models: *BAAI/bge-large-en-v1.5*[8] (Xiao et al., 2023) with 335M params, *thenlper/gte-large*[9] (Li et al., 2023) with 335M params, *Lajavaness/bilingual-embedding-base*[10] (Thakur et al., 2020) with 278M params and *intfloat/multilingual-e5-large-instruct*[11] (Wang et al., 2023) with 7.11B params. The results are summarized in Table 3.

Table 3: Comparison with pre-trained LLMs.

Model	ρ	p_ρ
BAAI/bge-large-en-v1.5	0.183	2.773E-28
thenlper/gte-large	0.193	2.207E-31
Lajavaness/bilingual-embedding-base	0.214	4.332E-38
intfloat/multilingual-e5-large-instruct	0.246	2.834E-50
best UK-trained model (our)	**0.269**	1.017E-55

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. Despite having fewer parameters and being trained on a smaller corpus, the domain-specific model demonstrated a stronger alignment with occupational relationships. This advantage can be attributed to the fact that the FastText model was trained directly on job advertisements, which inherently contain structured information about occupations, roles, and required skills, key factors that contribute to a

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

more accurate representation of occupational similarities.

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. If, for example, 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: $v_{size} = 300$, $\tau = 50$, $\mathbf{A}$ =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, generating a total of 100 different 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 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 measuring semantic consistency and the confidence interval of 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 (Malandri et al., 2024), confirming that many anchor sets generally lead to better alignment performance.

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.

Table 4: Best and worst NDCG threshold 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

5 Assessing Skill Bundles across Europe using VEUCTOR

We analyse the best and worst-aligned embeddings after training, aligning, and evaluating all models. The objective is to assess 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 the differences between the best and worst alignment models in 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 $S_o = \{s_1, s_2, ..., s_n\}$, 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. 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 when comparing the best and worst embeddings provides an indicator of the bias introduced by model selection.

As emphasized in the previous paragraphs, the development of embeddings in a multilingual setting poses significant challenges related to the selection of the best embeddings and their alignment for comparability purposes. Therefore, we start by performing an *intra*-country analysis, i.e. we compare the best and worst embeddings without aligning them. In this case, the analysis can be performed within countries, but the results cannot be compared between countries. Figure 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 Gold-

man Kaplan (Goldman and Kaplan, 2018) test rejects the equality of the two CDFs at 1%.[12] The region of rejection of the familywise error rate is highlighted by the thick black line underlying the graph. The right panel displays box-plots of the values of skill similarity by major occupation groups (1 Digit ESCO.). Also in this case, there is a significant difference in the distribution of skill similarity by occupation of the two models, with the best embeddings resulting in higher overall skill similarities across occupations.[13]

Our analysis has demonstrated 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. In this section, 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.

Our 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 embedding model proves to be of critical importance. Figure 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.

Finally, Figure 8 extends the cross-country analysis by comparing distributions across countries. The results demonstrate that the best embeddings consistently

[12] 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 to obtain the ranges of x (if it exists) where $F(x) = G(x)$ is rejected at the specified error level.

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

generate skill bundles that exhibit higher similarity across countries, reinforcing the robustness of this finding.

5.3 Reproducibility

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.[14]

Additionally, 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.

```
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[15] library is used to load the model. For the aligned model, the widely used gensim[16] library is employed to load and utilize the model. Examples are provided using Python to be in line 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.

Code Availability. All code used in this study was written in Python 3.12. The complete implementation is available in a GitLab repository (https://gitlab.com/crisp1/veuctor.git), which provides the scripts and additional data needed to reproduce our experiments.

[14] Due to the model sizes exceeding 150 GB, several Zenodo archives were created 10.5281/zenodo.15064201, 10.5281/zenodo.15064706, 10.5281/zenodo.15064719, 10.5281/zenodo.15064731, 10.5281/zenodo.15064738

[15] https://fasttext.cc/docs/en/python-module.html

[16] https://radimrehurek.com/gensim/

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 performs the evaluation of the aligned models. In addition, a demo script is provided to illustrate how to use both the models and the evaluation scripts. The repository also includes a README.md file with detailed instructions on use 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,[17] each equipped with 32 CPUs and 64 GB of memory, along with one high-performance machine featuring 64 CPUs and 128 GB of memory.[18] 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.

[17] AWS 3x c6a.8xlarge
[18] AWS c6a.16xlarge

6 Concluding Remarks

Every fisherman knows that the type of fish caught crucially depends on the bait which is used. In our context, not considering the possible bias generated by the choice of the embedding model is equivalent to inferring the composition of the fish population in a lake based on what is caught, failing to recognize that the type of bait heavily affects the catch.

In this paper, we have demonstrated that the choice of word embeddings significantly influences the type of information extracted from both structured and unstructured text. To address this, we 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.

A Appendix A

Algorithm 1 focuses on the generation and evaluation of word embedding models for different 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,

Algorithm 1 Embedding Pool Training and Evaluation (Step 1 & step 2)

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

$\theta_{\mathcal{C}}^{\top} \leftarrow \emptyset$
$\theta_{\mathcal{C}}^{\perp} \leftarrow \emptyset$
for $c \in \mathcal{C}$ **do**
    for $\theta \in \Theta$ **do**
        $\mathcal{M}_{c,\theta} \leftarrow \textit{train}(\mathcal{D}_c, \theta)$ ▷ Step 1. Train the model on corpus of country c
        $\mathbf{S}_{cos} \leftarrow \emptyset$ ▷ Step 2. Stores cosine similarities 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
    $\theta_{\mathcal{C}}^{\top} \leftarrow \theta_{\mathcal{C}}^{\top} \cup \theta_c^{\top}$
    $\theta_{\mathcal{C}}^{\perp} \leftarrow \theta_{\mathcal{C}}^{\perp} \cup \theta_c^{\perp}$
end for

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 how well the aligned vectors match the target model by calculating the cosine 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 emphasizes the need for 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.

Algorithm 2 Alignment Embedding Pool Training and Evaluation (Step 3 & step 4)

Require: $\mathcal{C}$ ▷ Set of countries analyzed
Require: Θ ▷ Set of hyperparameter configurations for embedding models
Require: $\{\mathcal{M}_c\}_{c\in\mathcal{C}}$ ▷ 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: $\theta_{\mathcal{C}}^{\top}$ ▷ Set of best parameters combination for each country
Ensure: $\theta_{\mathcal{C}}^{\perp}$ ▷ Set of worst parameters combination for each country

$\theta_{\mathcal{C}}^{\top} \leftarrow \emptyset$
$\theta_{\mathcal{C}}^{\perp} \leftarrow \emptyset$
for $c \in \mathcal{C}$ **do**
  for $\theta \in \Theta$ **do**
    $\mathcal{A}_{c,\theta} \leftarrow \mathit{SeNSe}(\mathcal{M}_c, \mathcal{M}_{target}, \theta)$ ▷ Step 3. Align the source model of c to the target model
    $\mathbf{S}_{cos} \leftarrow \emptyset$ ▷ Step 4. Stores cosine similarities scores
    for each $o \in \mathcal{O}$ **do**
      $v_{\mathcal{A}_{c,\theta}} \leftarrow \mathcal{A}_{c,\theta}(o)$ ▷ Encoder occupation o
      $v_{\mathcal{M}_{target}} \leftarrow \mathcal{M}_{target,\theta}(o)$
      $\mathbf{S}_{cos} \leftarrow \mathbf{S}_{cos} \cup cosine(v_{o_i}, v_{o_j})$
    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
  $\theta_{\mathcal{C}}^{\top} \leftarrow \theta_{\mathcal{C}}^{\top} \cup \theta_c^{\top}$
  $\theta_{\mathcal{C}}^{\perp} \leftarrow \theta_{\mathcal{C}}^{\perp} \cup \theta_c^{\perp}$
end for

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 demonstrate 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. (2020). 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.

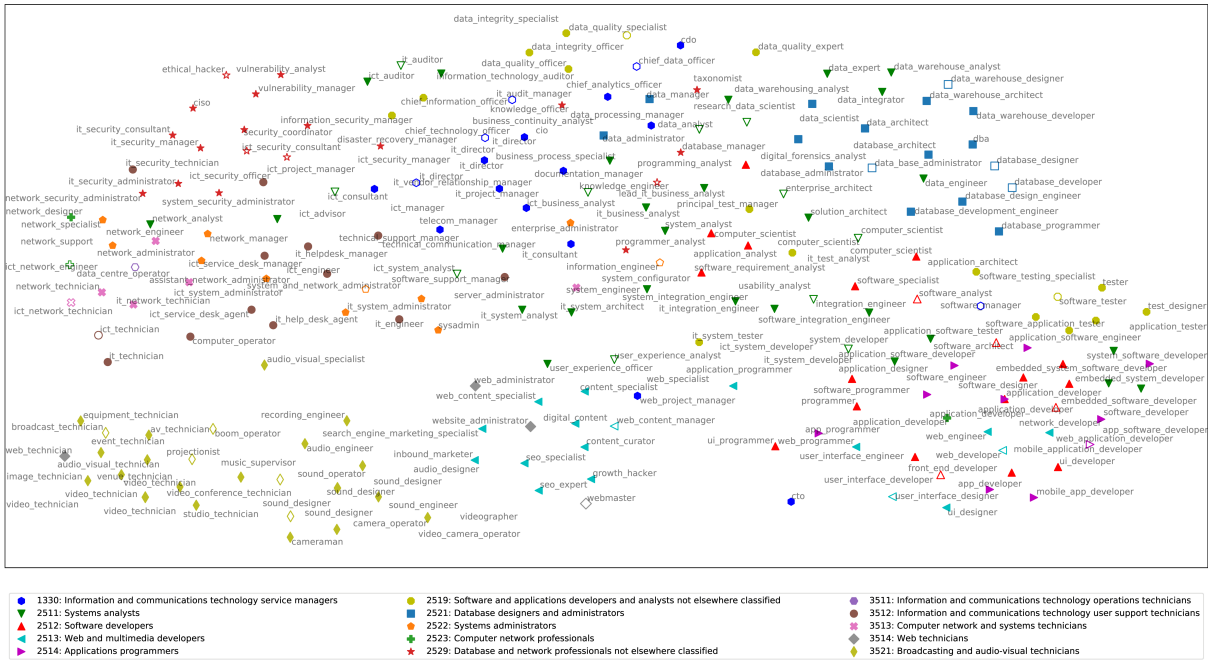


Figure 1: UMAP plot of the **best** word embedding model, according to HSS metric (Giabelli et al., 2020). The plot shows the ESCO concepts and words belonging to each group are shown distinguishing between narrower occupations (empty shape) and alternative labels (filled shape). Trained over 2 million ICT-related jobs in the UK. Taken from (Giabelli et al., 2020)

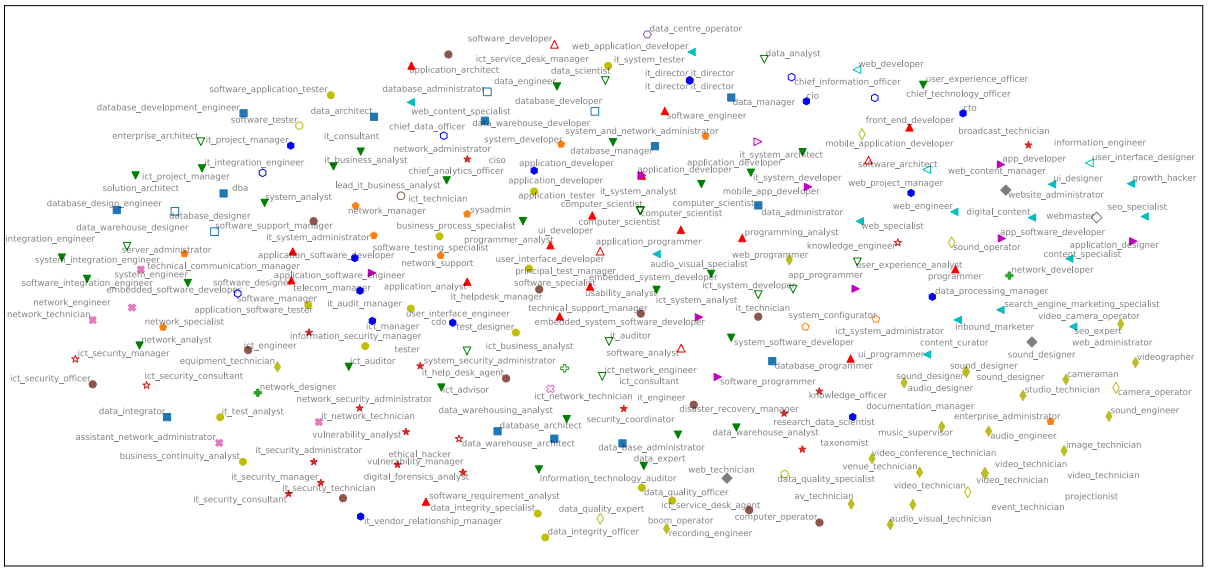


Figure 2: UMAP plot of the *worst* word embedding model, according to HSS metric (Giabelli et al., 2020). The plot shows the ESCO concepts and words belonging to each group are shown distinguishing between narrower occupations (shallow shape) and alternative labels (filled shape). Trained over 2 million ICT-related jobs in UK

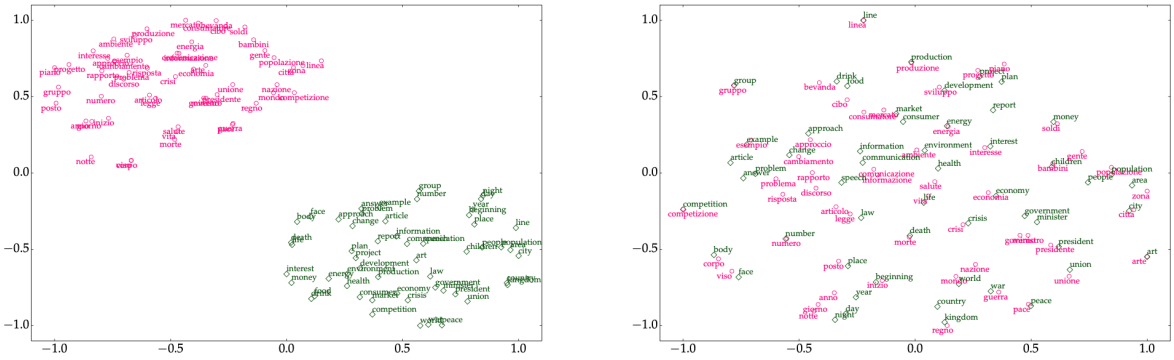


Figure 3: An image from (Ruder et al., 2019) depicting unaligned monolingual word embeddings (left) and word embeddings projected into a joint cross-lingual embedding space (right). Embeddings are visualized with t-SNE.

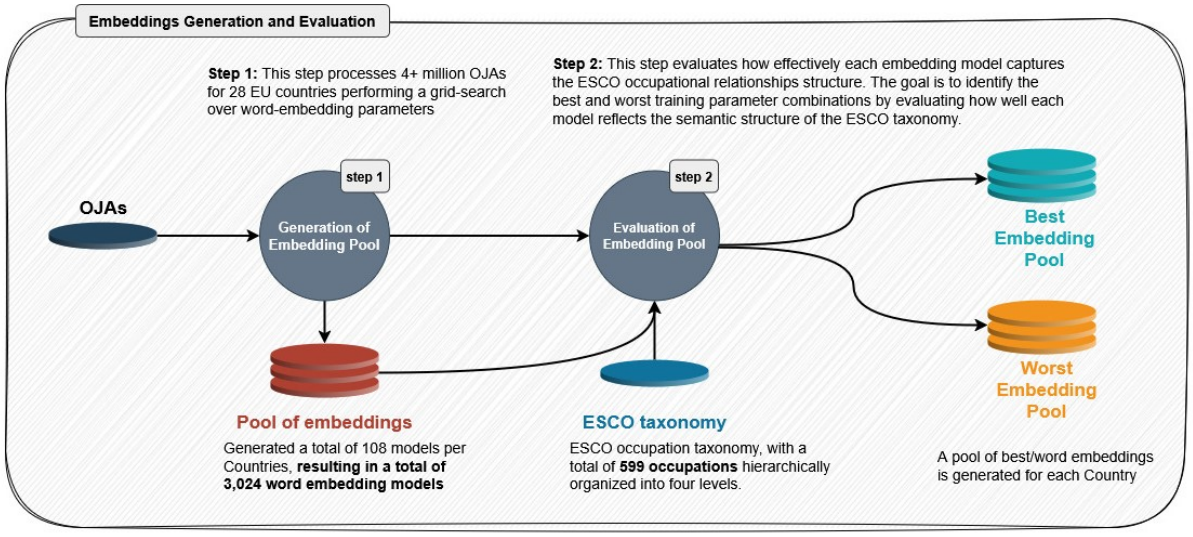


Figure 4: Workflow for the generation and evaluation of word embedding models.

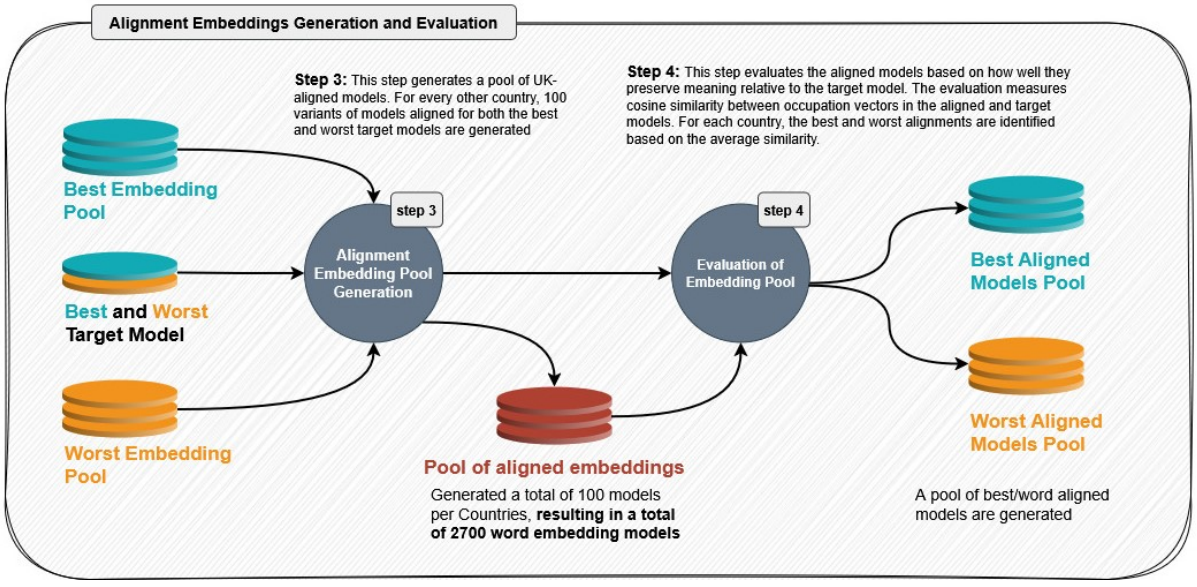


Figure 5: Workflow for the generation and evaluation of aligned models.

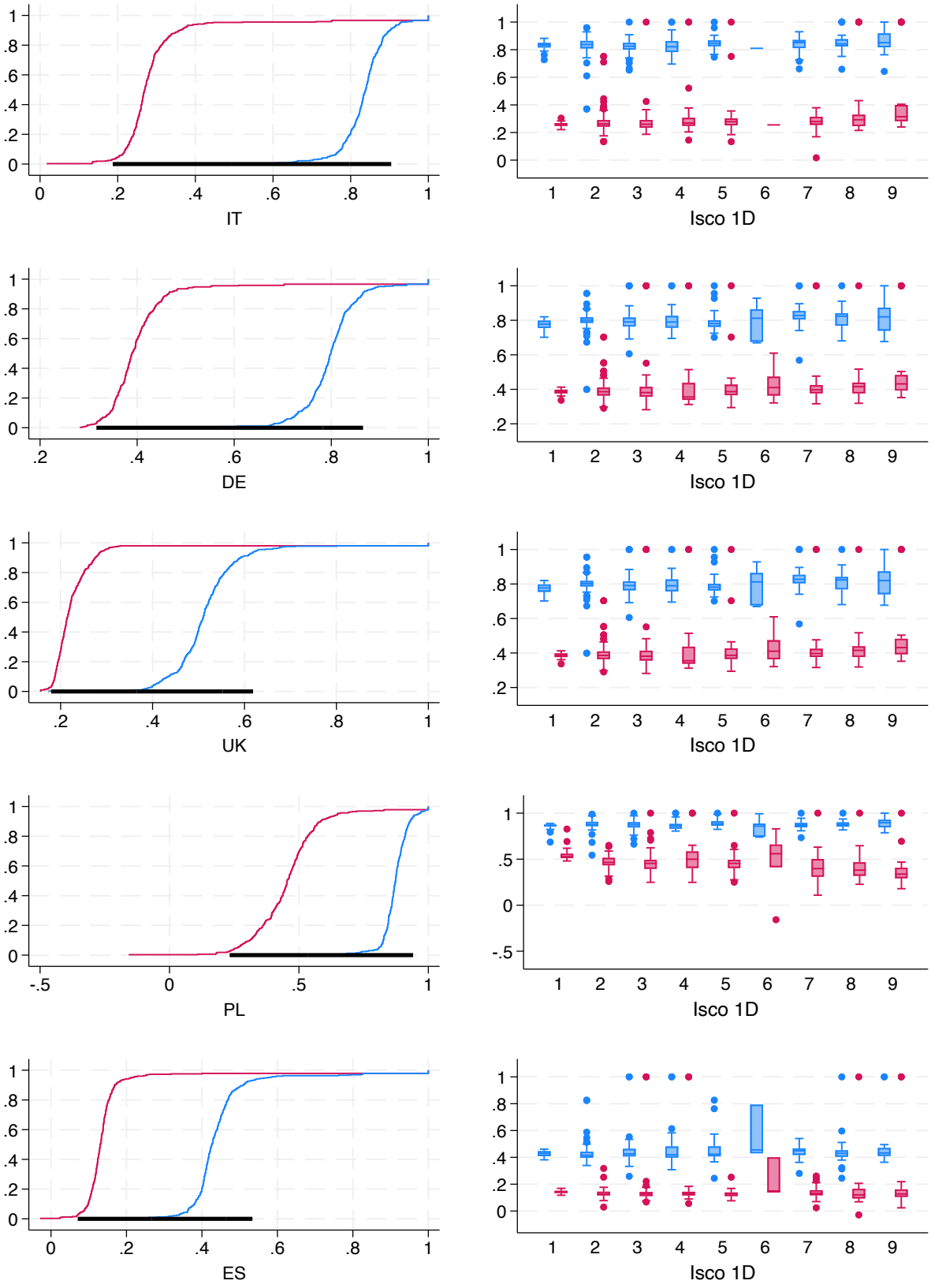
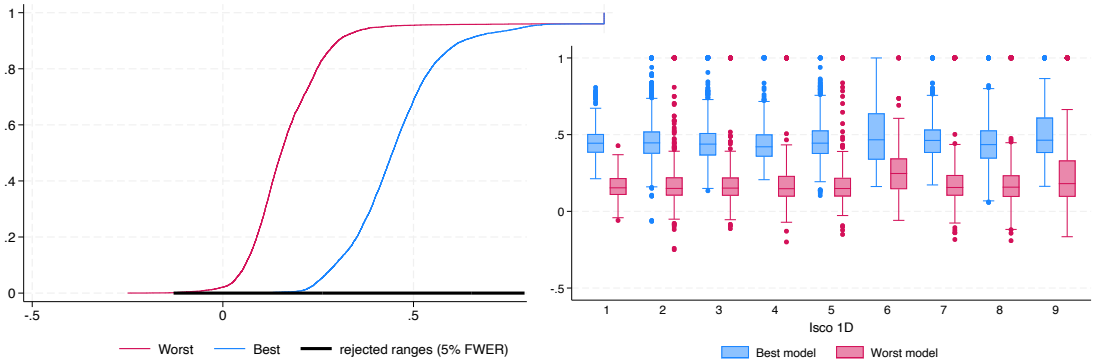


Figure 6: Distribution comparison: best and worst embeddings

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)



(a) Distribution comparison best and worst models

(b) Distribution of mean skill similarity over 1Digit ISCO

Figure 7: 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)

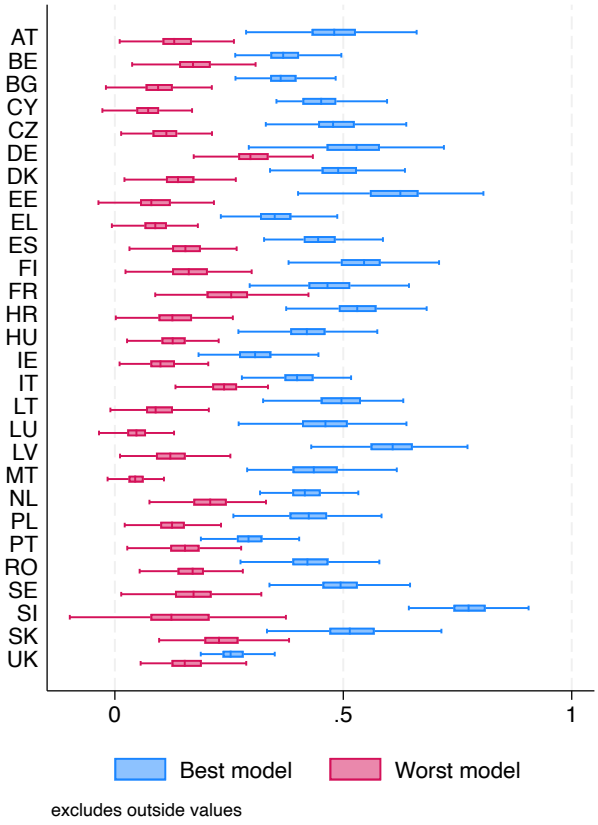


Figure 8: Distribution of mean skill similarity by country

References

Artetxe, M., Labaka, G., Agirre, E., 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. arXiv preprint arXiv:1805.06297 URL: https://arxiv.org/abs/1805.06297.

Azar, J., Huet-Vaughn, E., Marinescu, I., Taska, B., von Wachter, T., 2023. Minimum wage employment effects and labor market concentration. The Review of Economic Studies , rdad091doi:10.1093/restud/rdad091.

Azar, J., Marinescu, I., Steinbaum, M., Taska, B., 2020. Concentration in us labor markets: Evidence from online vacancy data. Labour Economics 66, 101886. doi:https://doi.org/10.1016/j.labeco.2020.101886.

Azpiazu, I.M., Pera, M.S., 2020. Hierarchical mapping for crosslingual word embedding alignment. Transactions of the Association for Computational Linguistics 8, 361–376.

Bhola, A., Halder, K., Prasad, A., Kan, M.Y., 2020. Retrieving skills from job descriptions: A language model based extreme multi-label classification framework, in: Proceedings of the 28th international conference on computational linguistics, pp. 5832–5842.

Bjerkander, L., Glas, A., 2024. Talking in a language that everyone can understand? clarity of speeches by the ecb executive board. Journal of International Money and Finance 149, 103200. doi:https://doi.org/10.1016/j.jimonfin.2024.103200.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146.

Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G., Viviani, M., 2018a. Wolmis: a labor market intelligence system for classifying web job vacancies. J. Intell. Inf. Syst. 51, 477–502.

Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., 2017. Using machine learning for labour market intelligence. ECML PKDD 2017: Machine Learning and Knowledge Discovery in Database , 330–342.

Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., 2018b. Classifying online job advertisements through machine learning. Future Generation Computer Systems 86, 319–328.

Braxton, J.C., Taska, B., 2023. Technological change and the consequences of job loss. American Economic Review 113, 279–316. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20210182, doi:10.1257/aer.20210182.

Chi, Z., Dong, L., Zheng, B., Huang, S., Mao, X.L., Huang, H.Y., Wei, F., 2021. Improving pretrained cross-lingual language models via self-labeled word alignment, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 3418–3430.

Clavié, B., Soulié, G., 2023. Large language models as batteries-included zero-shot esco skills matchers. arXiv preprint arXiv:2307.03539 .

Colace, F., Santo, M.D., Lombardi, M., Mercorio, F., Mezzanzanica, M., Pascale, F., 2019. Towards labour market intelligence through topic modelling, in: Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS), pp. 5256–5265. URL: http://hdl.handle.net/10125/59962.

Colombo, E., Marcato, A., 2023. Skill demand and labour market concentration: evidence from Italian vacancies. International Journal of Manpower 44, 156–198. doi:10.1108/IJM-04-2023-0181.

Colombo, E., Mercorio, F., Mezzanzanica, M., 2019. Ai meets labor market: Exploring the link between automation and skills. Information Economics and Policy 47. URL: http://www.sciencedirect.com/science/article/pii/S0167624518301318, doi:https://doi.org/10.1016/j.infoecopol.2019.05.003.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H., 2017. Word translation without parallel data. arXiv preprint arXiv:1710.04087 URL: https://arxiv.org/abs/1710.04087.

Decorte, J.J., Van Hautte, J., Demeester, T., Develder, C., 2021. Jobbert: Understanding job titles through skills. arXiv preprint arXiv:2109.09605 URL: https://arxiv.org/abs/2109.09605.

Decorte, J.J., Verlinden, S., Van Hautte, J., Deleu, J., Develder, C., Demeester, T., 2023. Extreme multi-label skill extraction training using large language models. arXiv preprint arXiv:2307.10778 URL: https://arxiv.org/abs/2307.10778.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 URL: https://arxiv.org/abs/1810.04805.

D'Amico, S., De Santo, A., Mercorio, F., Mezzanzanica, M., 2024. Enriching skill taxonomies through vector space models, in: 2024 IEEE International Conference on Big Data (BigData), IEEE. pp. 2297–2302.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., 2022. Weta: Automatic taxonomy alignment via word embeddings. Computers in Industry 138, 103626.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., Seveso, A., 2020. Neo: A tool for taxonomy enrichment with new emerging occupations, in: International Semantic Web Conference, Springer. pp. 568–584.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., Seveso, A., 2021a. Neo: A system for identifying new emerging occupation from job ads, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 16035–16037.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., Seveso, A., 2021b. Skills2job: A recommender system that encodes job offer embeddings on graph databases. Applied Soft Computing 101, 107049.

Godfray, H.C.J., 2002. Challenges for taxonomy. Nature 417, 17–19.

Goldfarb, A., Taska, B., Teodoridis, F., 2023. Could machine learning be a general purpose technology? a comparison of emerging technologies using data from online job postings. Research Policy 52, 104653. doi:https://doi.org/10.1016/j.respol.2022.104653.

Goldman, M., Kaplan, D.M., 2018. Comparing distributions by multiple testing across quantiles or cdf values. Journal of Econometrics 206, 143–166.

Gu, R., Zhong, L., 2023. Effects of stay-at-home orders on skill requirements in vacancy postings. Labour Economics 82, 102342. doi:https://doi.org/10.1016/j.labeco.2023.102342.

Hansen, S., McMahon, M., Prat, A., 2017. Transparency and deliberation within the fomc: A computational linguistics approach*. The Quarterly Journal of Economics 133, 801–870. doi:10.1093/qje/qjx045.

Harris, Z.S., 1954. Distributional structure. Word 10, 146–162.

Hershbein, B., Kahn, L.B., 2018. Do recessions accelerate routine-biased technological change? evidence from vacancy postings. American Economic Review 108, 1737–72. doi:10.1257/aer.20161570.

Javed, F., Hoang, P., Mahoney, T., McNair, M., 2017. Large-scale occupational skills normalization for online recruitment, in: Twenty-Ninth IAAI Conference.

Jiang, J.J., Conrath, D.W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008 .

Leacock, C., Chodorow, M., Miller, G.A., 1998. Using corpus statistics and wordnet relations for sense identification. Computational Linguistics 24, 147–165.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M., 2023. Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:2308.03281 URL: https://arxiv.org/abs/2308.03281.

Lin, D., et al., 1998. An information-theoretic definition of similarity., in: Icml, pp. 296–304.

Maedche, A., Staab, S., 2001. Ontology learning for the semantic web. IEEE Intelligent systems 16, 72–79.

Malandri, L., Mercorio, F., Mezzanzanica, M., Nobani, N., 2021. Taxoref: Embeddings evaluation for ai-driven taxonomy refinement, in: ECML PKDD.

Malandri, L., Mercorio, F., Mezzanzanica, M., Pallucchini, F., 2024. Sense: embedding alignment via semantic anchors selection. International Journal of Data Science and Analytics , 1–15.

Martin Henning, Rikard Eriksson, P.G.H.M., Elekes, Z., 2025. Job relatedness, local skill coherence and economic performance: a job postings approach. Regional Studies, Regional Science 12, 95–122. doi:10.1080/21681376.2025.2459148.

Mezzanzanica, M., Mercorio, F., 2019. Big data enables labor market intelligence, in: Sakr, S., Zomaya, A.Y. (Eds.), Encyclopedia of Big Data Technologies. Springer. URL: https://doi.org/10.1007/978-3-319-63962-8_276-1, doi:10.1007/978-3-319-63962-8_276-1.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 URL: https://arxiv.org/abs/1301.3781.

Mikolov, T., Le, Q.V., Sutskever, I., 2013b. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 URL: https://arxiv.org/abs/1309.4168.

Modestino, A.S., Shoag, D., Ballance, J., 2020. Upskilling: Do employers demand greater skill when workers are plentiful? The Review of Economics and Statistics 102, 793–805. doi:10.1162/rest_a_00835.

Muennighoff, N., Tazi, N., Magne, L., Reimers, N., 2022. Mteb: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316 URL: https://arxiv.org/abs/2210.07316, doi:10.48550/ARXIV.2210.07316.

Papoutsoglou, M., Ampatzoglou, A., Mittas, N., Angelis, L., 2019. Extracting knowledge from on-line sources for software engineering labor market: A mapping study. IEEE Access .

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training .

Resnik, P., 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of artificial intelligence research 11, 95–130.

Ruder, S., Vulić, I., Søgaard, A., 2019. A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research 65, 569–631.

Seco, N., Veale, T., Hayes, J., 2004. An intrinsic information content metric for semantic similarity in wordnet, in: Ecai, p. 1089.

Thakur, N., Reimers, N., Daxenberger, J., Gurevych, I., 2020. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. arXiv e-prints , arXiv–2010.

Turrell, A., Speigner, B., Djumalieva, J., Copple, D., Thurgood, J., 2018. Using job vacancies to understand the effects of labour market mismatch on uk output and productivity .

UK Commission for Employment and Skills, 2015. The importance of LMI, available at https://goo.gl/TtRwvS.

Vinel, M., Ryazanov, I., Botov, D., Nikolaev, I., 2019. 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. pp. 99–112.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F., 2023. Improving text embeddings with large language models. arXiv preprint arXiv:2401.00368 .

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., 2023. C-pack: Packaged resources to advance general chinese embedding. arXiv:2309.07597.

Zhang, M., Van Der Goot, R., Plank, B., 2023. Escoxlm-r: Multilingual taxonomy-driven pre-training for the job market domain. arXiv preprint arXiv:2305.12092 .