

# Harmonizing occupational classifications in the UK health sector: challenges and opportunities for SOC 2020 and ISCO-08 integration

Digital  
Transformation  
and Society

Sonia Elizabeth Herrera-Salgado  
*University of Stirling Business School, Stirling, UK*

Received 25 August 2025  
Revised 5 November 2025  
19 December 2025  
16 February 2026  
21 March 2026  
Accepted 28 March 2026

## Abstract

**Purpose** – Occupational classification (OC) systems like SOC 2020 and ISCO-08 are vital for healthcare workforce planning; however, differences between these standards reduce comparability. This study examines the technical feasibility of harmonizing UK health OC through natural language processing (NLP) models.

**Design/methodology/approach** – First, a hierarchical analysis of OC systems was developed. Then, NLP models (i.e. Bag of Words, TF-IDF, SBERT and an ensemble) were proposed for semantic mapping and benchmarking of OC systems in the United Kingdom. A manual validation study with two domain-aware annotators assessed mapping accuracy.

**Findings** – We found that the SBERT model matches occupations more accurately than other models. Indeed, 1,308 matches from one job title to many job titles pairings were identified. Moreover, we benchmarked four UK health systems and showed that using SBERT, OC systems can achieve an average similarity and standard deviation of 63.21% (17.13%) for Scotland-Wales, 58.80% (16.49%) for Scotland-England and 61.28% (15.71%) for Scotland-Northern Ireland. Manual validation yielded precision scores of 0.73–0.98 and F1 scores of 0.84–0.99 across datasets. Using NLP simplifies the classification process, reduces administrative tasks and improves consistency with international standards. However, agreement levels varied across comparisons (76–90%), indicating that expert oversight remains necessary.

**Originality/value** – This study shows the technical feasibility of using NLP models for healthcare OC harmonization and offers a practical and scalable solution for improving OC alignment. This research applies NLP models jointly with benchmarking processes prioritizing practical adoption over methodological novelty, focusing on the UK healthcare sector.

**Keywords** Occupational classification system, Healthcare workforce, Semantic mapping, Granularity, NLP, SBERT

**Paper type** Research article

## Introduction

Occupational titles are usually categorized according to the International Standard Classification of Occupations (ISCO), which is the globally accepted system by the International Labour Organization (ILO) (ILOSTAT-1, 2024). Most National Statistical Offices (NSOs) have adopted the ISCO-08 classification, either as their primary system or in conjunction with existing national classification systems (Tijdens & Kaandorp, 2018). The Standard Occupational Classification (SOC) 2020 is the UK's current classification system. It is widely used by employment agencies for job matching and for labour market policy (ONS, 2024). Both systems share a four-level hierarchical structure (major, sub-major, minor, and unit groups), though ISCO-08 contains 436 occupational units compared to SOC 2020's 412. SOC 2020 uses the same principles as ISCO-08; however, aligning these systems at finer granularity levels is challenging because a single unit group in one classification may map to multiple unit groups in the other (Schmidtke, 2014).

© Sonia Elizabeth Herrera-Salgado. Published in *Digital Transformation and Society*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at [Link to the terms of the CC BY 4.0 licence](#).

*Conflict of interest:* I have no conflicts of interest to disclose.



Digital Transformation and Society  
Emerald Publishing Limited  
e-ISSN: 2755-077X  
p-ISSN: 2755-0761  
DOI 10.1108/DTS-06-2025-0270

Despite the widespread adoption of ISCO-08 and SOC 2020 as foundational frameworks for OC, existing research and practice reveal significant challenges in harmonizing these systems, particularly in the healthcare sector, where job roles and skill requirements are complex and continuously evolving. This fragmentation limits the comparability of workforce data across national and international contexts, thereby constraining effective workforce planning and international benchmarking.

In this study, we comprehensively analyse and compare OC systems within healthcare by examining data from the ILO alongside the UK's National Health Service classifications. By employing hierarchical analysis, semantic mapping, and a benchmarking process, we aim to identify inconsistencies and barriers in the harmonization between these systems. Additionally, we use NLP models for the semi-automatic classification processes, proposing a scalable solution that enhances alignment with global standards while maintaining local relevance. This study demonstrates technical feasibility rather than deployment-ready solutions.

This article is structured as follows: First, we provide background on existing methods. Next, we describe our cascade method comprising Hierarchical Analysis, Semantic Mapping, and Benchmarking. Then, results of comparative analysis are shown. Finally, the feasibility and practical implications of adopting SOC in the NHS.

## Background

A study developed by the Warwick Institute of Employment Research identified how SOC 2020 and ISCO-08 could be matched at the job title level (Warwick, 2018). They created a coding tool called CASCOT (Computer Assisted Structured Tool), which took SOC job titles as input and produced ISCO codes as output. In some cases, there was not a perfect match, leaving some titles without a match. Beyond CASCOT, recent tools include SOCcer for US SOC codes (Russ *et al.*, 2016, 2023); AUTONOC for Canadian NOC (Suarez Garcia *et al.*, 2021); Labour for ISCO (Kouretsis, Bampouris, Morfiris, & Papageorgiou, 2020) and transformer-based models (Safikhani, Avetisyan, Föste-Eggers, & Broneske, 2023). A 2023 comparison found that existing tools achieved only 17–26% exact agreement at the 4-digit level (Ge, Friesen, Locke, Russ, Burstyn, Baker, . . . , & Huss, 2023), highlighting the need for more accurate methods. Young, Fedkina, Chatwood, and Bjerregaard (2018) conducted a study on the healthcare workforce in eight Arctic states, using a semantic and linguistic approach due to the complexity of extracting descriptive information at a high level of granularity. This methodology enabled comparability between the eight states and helped the researchers identify trends and patterns within workforce job titles despite the challenge of working with different languages. Giunchiglia and Shvaiko (2003) used an approach called semantic matching by mapping concepts between two sources of qualitative data at two levels of granularity: element and structure. Zhang, Ren, and Li (2017) proposed an ontology integration method (OIM) defined by two sub-processes: the mapping process and the integration process (merging the mapped concepts). It takes two ontologies as input, searches for correspondences and produces a set of semantic mappings. Semantic web research has proposed a solution for the recovery of precise information by ontology matching techniques and intelligent agent technology (Rana & Singh, 2014). Cheong, Yin, Cheung, Fung, and Poon (2023) proposed a representation-learning framework to adapt the structure of medical ontologies for robust integration with electronic health records based on NLP models. A framework for weakly supervised entity classification using medical ontologies and expert-generated rules was presented to analyse the records of patients (Fries *et al.*, 2021). While SNOMED-CT and FHIR provide clinical terminology standards, occupational classifications serve distinct workforce planning purposes. Shahzad *et al.* (2021) proposed an ontological framework for smart healthcare services that incorporated similarities, differences, dependencies and other semantic relations.

Mannetje and Kromhout (2003) suggest that the selection of an OC system for healthcare research should be guided by the following key criteria: (1) the presence of a well-defined

---

hierarchical structure, (2) a complete set of job titles across multiple languages and (3) adherence to standardized frameworks that facilitate flexibility and comparability across datasets and studies. Hierarchical classifications were used by [Kuodytė and Petkevičius \(2021\)](#) to predict occupations from one or many completed educational programs.

[Krasna, Venkataraman, Robins, Patino, and Leider \(2024\)](#) conducted a study to ascertain the feasibility of implementing the Department of Labor’s SOC codes within the public health workforce in the United States. The study adopted a benchmarking approach to gain insight into the categorization of healthcare job titles. The occupational taxonomy and semantic similarities enabled unique job titles to be categorized. [Abraham et al. \(2023\)](#) analysed job title similarity using occupational information derived from network data, showing that robust benchmarking requires assessment across job family design, occupational classification and job evaluation systems to support more harmonized occupational data frameworks for policy and research.

In Scotland, NHS workforce statistics, as reported by [Turas \(2024\)](#), categorize healthcare occupations into two primary groups: medical and dental, and non-medical and dental. In England, NHS workforce data for Hospital and Community Health Services employs a hierarchical coding system composed of five groups with varying levels of detail ([NHS England, 2024](#)). Northern Ireland’s [Department of Health \(2024\)](#) classifies healthcare occupations using two levels: Main Staff Group and Job Family. In Wales, according to [StatsWales \(2024\)](#), health and social care staff are divided into two main groups: non-medical staff, and medical and dental staff.

Together, these coding structures across the UK’s four nations reveal a shared emphasis on hierarchical classification, balancing aggregation for broad analysis with detailed job-level granularity to capture the diversity of healthcare roles. Differences in grouping and coding reflect regional administrative practices and confidentiality considerations.

[Susan, Sharma, and Choudhary \(2024\)](#) proposed to use NLP models for matching resumes to job profiles and found that including embeddings performs better than a standalone bag of words approach. [Yilmaz et al. \(2022\)](#) explored NLP and machine learning (ML) techniques for identifying trainees at risk and demonstrated that qualitative data require supervisory input to organize and interpret.

Bag of Words (BoW) was used by [Cichosz \(2023\)](#) for medical article classification. BoW proved to be effective for tasks involving short text classification ([Dai et al., 2024](#)). [Li et al. \(2007\)](#) applied TF-IDF approach for the extraction of Chinese texts analysing linguistic characteristics of documents, including uni/bi/trigram extraction. TF-IDF balances local word importance within a document and global rarity across the corpus ([Guleria, Frmda, & Naga Srinivasu, 2025](#)). The SBERT model is used to provide semantic information from oral production samples of elderly controls and people with Alzheimer’s disease ([Dhawaleswar Rao & Pani, 2025](#); [Santander-Cruz, Salazar-Colores, Paredes-García, Guendulain-Arenas, & Tovar-Arriaga, 2022](#)).

The semantic matching successes of [Susan et al. \(2024\)](#), [Cheong et al. \(2023\)](#) and [Fries et al. \(2021\)](#) in healthcare contexts motivated our selection of SBERT over traditional approaches. The benchmarking frameworks of [Abraham et al. \(2023\)](#) and [Krasna et al. \(2024\)](#) provided templates for our UK system comparisons. OIMs ([Zhang et al., 2017](#); [Shahzad et al., 2021](#)) guided our hierarchical analysis approach. Specifically, CASCOT’s hierarchy informed our granularity matching; [Ge et al. \(2023\)](#) 17–26% lexical agreement motivated SBERT adoption; [Russ et al. \(2016\)](#) SOCcer’s v2 agreement 50%, established our threshold below 30%; [Kouretsis et al. \(2020\)](#) confidence bands set 60–80% review levels; and [Susan et al. \(2024\)](#) semantic-heavy ensemble approach guided our weighting.

## Research methodology

The literature review shows the necessity to implement a hybrid methodology to address the problem of OC systems. Moreover, there are no studies developed for healthcare systems in the United Kingdom. Here, we propose a methodology called the “cascade method” that consists of

three stages: hierarchical analysis, semantic mapping and benchmarking analysis performed to compare SOC and ISCO-08, and UK OC systems. For semantic mapping, the study explores four NLP models: BoW, TF-IDF, semantic embedding and an ensemble of the three.

#### *Data processing*

All data were acquired concurrently, with a uniform cut-off date of December 2023, and are publicly available at GitHub repository <https://github.com/xxxxx/Data-Science-Projects>. This enabled the benchmarking of the four systems by country during the same period.

The data processing focussed on pre-processing texts according to the occupational descriptions. The steps for the data cleaning are described as follows:

- (1) NA, NAN cases or text without description are replaced with a blank space “ ”.
- (2) Lowercasing, tokenization, stop words removal, stemming/lemmatization ([Table 1 – supplementary material 1](#)).
- (3) Clean special characters like: “^”, “w”, “s”, “-”, “&” keeping letters, numbers and some relevant symbols.
- (4) Normalize spaces using regular expressions patterns.

Once the data were pre-processed, the study of the hierarchical structure started. [Table 1](#) shows that SOC 2020 and ISCO-08 have the same granularity level and hierarchical structure, with the difference that ISCO-08 has a greater number of job titles.

Moreover, the number of job titles compared among the four UK OC systems shows a high level of heterogeneity by level of granularity and hierarchical structure. Some systems have divisions while others do not, as shown in [Table 2](#):

Once the data is hierarchized the next step is to select the best NLP model for the semantic mapping stage.

#### *NLP models*

BoW is one of the most common methods for text classification. In the BoW method, features are directly related to word occurrence frequencies. Formally, text like sentences, paragraphs or documents are represented by a real vector  $x = (x_1, x_2, \dots, x_n) \in R^n$ , where  $x_i$  is the frequency of a word  $w_i$  from a vocabulary  $W = (w_1, w_2, \dots, w_n)$  appearing in the text.

The term frequency-inverse document frequency (TF-IDF) is a variant of BoW. It combines two components: i) term frequency (TF) where for each term the value of the equivalent feature is the number of the term’s occurrences in a given document ([Cichosz, 2023](#),

**Table 1.** Summary comparison between ISCO-08 and SOC 2020 standards of OC, classified by four granularity levels, based on its homogeneous hierarchical structure from major until unit, and their corresponding number of job titles, with the most granular level being the “unit”

Standards occupational classification	Granularity	Hierarchical structure	Number job titles
ISCO – 08	4	Major	10
		Sub major	43
		Minor	130
		Unit	436
SOC 2020	4	Major	9
		Sub major	26
		Minor	104
		Unit	412

**Table 2.** Summary comparison of UK OC systems classified by granularity levels, based on their heterogeneous hierarchical structure, division of some systems and number of job titles

UK countries	Granularity	Hierarchical structure	Division	Number job titles
Scotland	2	Job family	Medical and Dental	72
		Sub Job family	Non-Medical and Dental	84
England	5	Main staff group	N/A	457
		Staff group 1		
		Staff group 2		
		Care Settings		
		Level		
North Ireland	2	Main staff group	N/A	626
Wales	3	Job family		
		First level	Medical and Dental	76
		Second level	Other Non-Medical and Dental	82
		Third level	Nursing	111

p. 606), and ii) inverse document frequencies (IDF) to normalize these terms frequencies, reducing the weight of repeated terms and increasing the weight of unique or less common words. The overall approach works through the next procedure, given a document collection  $D$ , a word  $w$ , and an individual document  $d \in D$ :

$$w_d = f_{w,d} * \log\left(\frac{|D|}{f_{w,D}}\right),$$

where  $f_{w,d}$  is the number of times  $w$  appears in  $d$ ,  $|D|$  is the size of the corpus and  $f_{w,D}$  is the number of documents in which  $w$  appears in  $D$  (Jalilifard, Caridá, Mansano, Cristo, & da Fonseca, 2021).

Although BoW and TF-IDF are widely used, they have difficulty capturing contextual relationships between words (Abubakar, Umar, & Bakale, 2022). To overcome this drawback, Graff, Moctezuma, and Téllez (2025) proposed using BERT (Devlin, Chang, Lee, & Toutanova, 2019) and Transformer-based models (Ahuir, Hurtado, García-Granada, & Sanchis, 2023). However, a key limitation of the native BERT architecture is its inefficiency for sentence-pair regression tasks due to its cross-encoder design. Therefore, we used SBERT introduced by Reimers and Gurevych (2019).

SBERT is a modification of the pretrained BERT network that leverages Siamese and triplet network architectures to derive semantically meaningful sentence embeddings. SBERT appends a pooling operation, typically computing the mean of all output vectors, to the output of BERT or RoBERTa to generate fixed-sized sentence representations. For sentence pairs, the classification objective function concatenates the sentence embeddings  $u$  and  $v$  with their element-wise difference  $|u - v|$  and processes them through a softmax classifier. Table 3 provides a synopsis of the objectives associated with the four methods.

**Table 3.** Objectives by NLP methods applied within the methodological framework “cascade method”, to understand the main differences between them

Method	Objective
BoW	To identify topics with simple frequencies
TF-IDF	To place greater weight on rare words
SBERT	To derive semantically meaningful sentence embedding
Ensemble	To improve model accuracy, robustness, and generalizability by combining the weights of three methods

### Analytical procedures

The methodology has two goals: a comparative analysis of SOC 2020 and ISCO-08 and of UK OC systems. The steps of the hybrid method for the first and second purposes are detailed below:

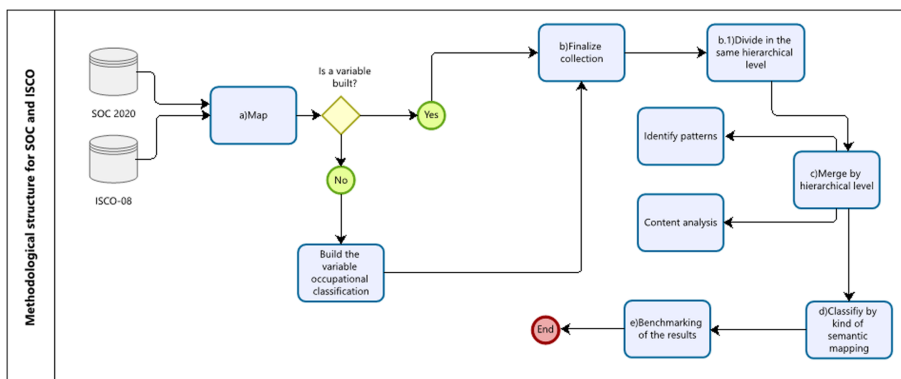
- (1) Map the SOC 2020 and ISCO-08 OC coding.
- (2) Finalize collection: formatting uniformly and building a variable with the job titles of the SOC 2020 and ISCO-08.
  - Divide the variable by hierarchical level.
- (3) Merge by hierarchical level the job titles using a semantic comparison approach.
- (4) Classify by kind of semantic matching, to measure the level of match.
- (5) Benchmarking of the results obtained.

The second method must first identify whether a hierarchical level exist. Figures 1 and 2 show the detailed workflow of the study.

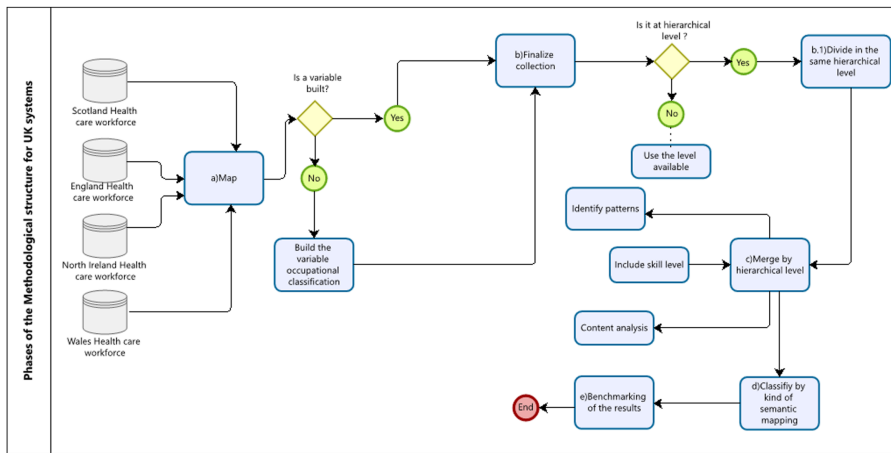
All tasks, including data pre-processing, and model implementation, were performed using the Python programming language and its associated data science libraries (Python Software Foundation, 2023). The libraries *pandas* v2.3.3 and *NumPy* v2.3.0 were used for data manipulation. Furthermore, the following libraries for NLP were used: *SentenceTransformer* v5.1 to generate dense vector embeddings for sentences, paragraphs or even entire documents; *cosine\_similarity* and *TfidfVectorizer* from *scikit-learn* v1.7.2.

Next, we applied NLP models to the datasets as follows:

- (1) The first step is to choose the type of semantic model: BoW, TF-IDF, SBERT or ensemble. The SBERT implementation uses paraphrase-multilingual-MiniLM-L12-v2 model; pooling: mean; max tokens: 128; batch size: 32; and random seed: 42.
- (2) To calculate the similarity using BoW. First, we combined all the texts to create the vocabulary. Then, for every document in the vocabulary, a vector is assigned, each dimension equivalent to a keyword and its frequency in the related document (Lwin et al., 2025, p. 45).
- (3) To vectorize both Excel files that will be compared and separate the matrices bow1 and bow2.



**Figure 1.** Diagram of the phases for methodological structure applied at the SOC and ISCO comparison. Beginning with the mapping of both sources to identify the hierarchical level, to make a match using NLP algorithms and to get an analysis of the results by benchmarking technique



**Figure 2.** Diagram of the phases for the methodological structure applied at the four UK OC systems. Beginning with the mapping of the four sources to identify the hierarchical level, to make a match using NLP algorithms and to get an analysis of the results by benchmarking technique

- (4) To calculate the cosine similarity of both matrixes and store in the matrix. The cosine similarity is a metric that measures the similarity between two vectors in an  $n$ -dimensional space (Prasad *et al.*, 2025). It is defined as follows:

$$\text{cosine}_{similarity}(P, Q) = \frac{\sum_{i=1}^n P_i Q_i}{\sqrt{\sum_{i=1}^n P_i^2} \sqrt{\sum_{i=1}^n Q_i^2}},$$

where  $P$  and  $Q$  are real vectors. The values returned are between  $-1$  and  $1$ , where  $1$  “identical vectors”,  $0$  “completely different”, and  $-1$  “opposite vectors”.

- (5) To calculate the similarity using the TF-IDF model. All the texts were combined to create the vocabulary. The next step was to vectorize using TF-IDF, with optimization for English words. We then proceeded to separate the matrices of both text files, calculate the cosine similarity, and store the results in a matrix.
- (6) To calculate the semantic similarity, we used a function that allows us to include the contextual information: i) encoding the text (vector), ii) creating the semantic relationships and correlation between words through the cosine similarity and iii) getting the embedding through the results of the similarity matrix.
- (7) To find the best matches between both tables of occupations, use the function *find\_best\_matches*, where the arguments are: i) both data frames, ii) both columns with occupations descriptions, iii) both identity numbers, iv) margin or threshold, value between  $0-1$ , v) top\_n = number of best matches for occupation and vi) methods of similarity to use: BoW “bow”, TF-IDF “tfidf”, SBERT “semantic”, and combine BoW (25%), TF-IDF (25%) and semantic (50%) “all” (ensemble), with weights justified by ablation analysis comparing eight configurations (Table 2 – supplementary material 1). Semantic-heavy configurations consistently outperformed traditional lexical approaches across all datasets. Optimal ensemble weights were 0.2/0.3/0.5 (BoW/TF-IDF/SBERT); threshold sensitivity analysis

showed 30% minimizes false rejections while 50% maximizes precision; SBERT alone achieved 89% of ensemble performance.

- (8) To generate a statistical report of the similarities, which includes: total matches, average similarity, standard deviation of similarity, high confidence matches with a similarity score greater than 80%, medium confidence matches in the range between 60% and 80%; and low confidence matches with a similarity score less than 60% and greater than 30%. The 30% threshold was selected based on sensitivity analysis evaluating 30%, 40% and 50% thresholds across all datasets ([Table 3 - supplementary material 1](#)). Higher thresholds may be preferred for production systems requiring greater precision.
- (9) Finally, to assess inter-rater reliability, 100 mappings were randomly sampled per dataset, stratified across similarity bands. Two domain-aware annotators independently labelled matches as correct, partially acceptable or incorrect. Labelling criteria defined: (1) Correct: equivalent job role; (2) Partially acceptable: related but not identical; (3) Incorrect: unrelated occupations. For metric calculations, “partially acceptable” mappings were counted as correct (true positives), as these represent functionally usable matches requiring only minor contextual adjustment. Agreement rates, Cohen’s Kappa, precision, recall and F1 scores are reported in ([Table 4](#)).

## Results and discussion

### *Comparative analysis of NLP models*

Cases with similarity scores above 80 are classified as high confidence, those scoring 60–80 as medium confidence and those below 60 as low confidence. The results of the match between SOC 2020 and ISCO-08 at the unit level using BoW, TF-IDF, SBERT and ensemble, show that standard deviation of TF-IDF was greater than the others. The second most consistent method is the ensemble; it presents good stability. SBERT demonstrated the most stable and reliable performance with the highest average similarity (68.39%) and lowest standard deviation (13.25%), identifying 1,308 matches compared to 421 for BoW, 264 for TF-IDF and 1,096 for the ensemble ([Table 5](#)).

[Table 5](#) shows that the group with medium confidence contains 65 job titles that match the BoW approach, TF-IDF has less matches with 31 matches, but the SBERT model can find 713 matches. Note that the ensemble model can identify 936 matches for the low confidence group; this is almost three times as many as those identified by SBERT.

The difference between groups with low confidence and medium confidence is the level of granularity of job titles within the SOC system. For example, in the ISCO group, non-commissioned Armed Forces Officers and Armed Forced Occupations, Other ranks are grouped in two codes, whereas in the SOC system, the algorithm identifies one code related to officers in the armed forces.

**Table 4.** Manual validation metrics of 100 sample mappings for the semantic embedding method (SBERT) across four UK OC systems, where two annotators participated

Dataset	Precision	Recall	F1	Accuracy	Agreement	Kappa
SOC-ISCO	0.88	1.00	0.94	0.88	76%	0.23
Scotland-N.Ireland	0.73	1.00	0.84	0.73	79%	0.54
Scotland-Wales	0.83	1.00	0.91	0.83	88%	0.63
Scotland-England	0.98	1.00	0.99	0.98	90%	0.67

**Table 5.** Descriptive measures: total number of matches, average similarity, standard deviation, and number of matches by high, medium and low confidence similarity of the comparison at unit level of SOC and ISCO are detailed through the methods BoW, TF-IDF, SBERT and Ensemble

	BoW	TF-IDF	SBERT	Ensemble
Matches found	421	264	1,308	1,096
Average similarity	50.22%	52.36%	68.39%	47.75%
Standard deviation	19.27%	22.45%	13.25%	14.78%
High confidence	38	39	238	52
Medium confidence	65	31	713	108
Low confidence	318	194	357	936

Table 6 shows that the highest similarity average between Scottish and English OC systems is 58.80% for the SBERT method, but the method with the least variability is the ensemble. For Scotland-Wales (84 vs. 82 job titles in non-medical categories), SBERT achieved 236 matches with 63.21% average similarity. For Scotland-England (84 vs. 61 job titles), SBERT achieved 229 matches. For Scotland-Northern Ireland (84 vs. 626 job titles), SBERT achieved 252 matches despite the significant size difference in job universes. At the level of the NHS England system, BoW yields the fewest matches, whereas SBERT yields the most.

Table 6 shows that even though the method with greater matches (252) is SBERT, TF-IDF shows the least variability. Although the Irish system's nomenclature includes an alphanumeric code at the beginning of every job title description, the SBERT method can identify 107 with low confidence similarity. Pattern analysis shows that higher similarity scores occur in occupation categories with standardized terminology (e.g. nursing roles), while lower scores appear in administrative and support roles with region-specific naming conventions. The Northern Ireland system's alphanumeric codes do not impair SBERT's performance, demonstrating its robustness to formatting variations.

**Table 6.** Descriptive measures: total number of matches, average similarity and standard deviation, and number matches by high medium and low confidence similarity of the comparison of non-medical and dental job titles of Scottish and Welsh systems, Sub Job family and Level job titles of Scottish and English system, and Sub Job family and Job titles of Scottish and North Irish system, are detailed through the methods BoW, TF-IDF, SBERT and Ensemble

OC systems	Matches	Average similarity	Standard deviation	High confidence	Medium confidence	Low confidence
<i>BoW</i>						
Scotland-Wales	63	57.29%	24.70%	13	2	48
Scotland-England	38	47.88%	18.85%	3	1	34
Scotland-North Ireland	116	46.13%	18.71%	5	19	92
<i>TF-IDF</i>						
Scotland-Wales	47	60.73%	26.21%	13	2	32
Scotland-England	22	53.68%	20.46%	3	0	19
Scotland-North Ireland	38	55.46%	9.09%	0	12	26
<i>SBERT</i>						
Scotland-Wales	236	63.21%	17.13%	41	92	103
Scotland-England	229	58.80%	16.49%	29	76	124
Scotland-North Ireland	252	61.28%	15.71%	29	116	107
<i>Ensemble</i>						
Scotland-Wales	159	49.15%	19.30%	14	22	123
Scotland-England	113	46.44%	14.25%	3	11	99
Scotland-North Ireland	192	46.70%	12.29%	2	27	163

The similarity of low confidence collects more matches in the ensemble method through all the comparisons. In the rest of the cases, the SBERT is the model with the best performance and exhibits minimal runtime (Table 4 – supplementary material 1). It was observed that both methods BoW and TF-IDF make some inaccuracies, i.e. the matching among the job title IT managers of NHS of Scotland with different managers of Wales as Education, Mining and Finance.

The validation sample was stratified across similarity bands to ensure high-risk mapping representation. Sample size aligns with comparable tools (SOCcer, AUTONOC). Expanded validation is recommended before deployment. Error analysis identified three types: (1) Granularity mismatch: specialist roles matched to generalist categories; (2) Semantic overlap: similar terminology, different functions; (3) Regional terminology: region-specific naming conventions. Clinical roles achieved 88% mapping accuracy versus 77% for non-clinical. By category: support workers 100%, nursing 93%, management 90%, AHPs 89%, estates 67%. Specialist mis-mappings comprised 47% of errors (e.g. technicians mapped to wrong domains), followed by regional terminology gaps (34%) (Table 5 – supplementary material 1). The recall of 1.00 is an artefact of the validation sampling strategy: samples were drawn from proposed matches only (no rejected pairs = no false negatives). The SOC-ISCO Kappa of 0.23 reflects cross-system ambiguity; higher values (0.54–0.67) for within-UK comparisons indicate better consistency.

Finally, we present a worked example for the validation procedure (Table 7):

#### *Benchmarking analysis*

The four hierarchical levels within the SOC 2020 and ISCO-08 are as follows: Major Group, Sub-major, Minor and Unit. The relationship between the Major Group in SOC 2020 and the Major Level in ISCO-08 shows that, at a lower level of job title granularity, it becomes more difficult to determine precisely which job title in one classification corresponds to its counterpart in another. However, the SBERT model identified 1,308 matches with an average similarity of 68.39%.

The non-medical and dental classification system of the NHS in Wales has three levels of granularity, whereas the classification system of the NHS in Scotland is grouped into two levels called “job families” and “sub-job families”. Comparing the job titles of the non-medical and dental group of the NHS of Scotland’s Sub job family with the third level of classification of Other Non-medical staff in the NHS of Wales has allowed us to identify a greater number of matches for the SBERT approach, against the rest of the models as shown in Table 6.

A comparison of the job family of the non-medical and dental group in the NHS of Scotland with the main staff group in the NHS of England classification system reveals a significant heterogeneity of granularity. To achieve a suitable level of comparability between the two systems, a comparison has been made at the level of the sub-job family in the NHS of Scotland with the fourth level in the NHS of England. Nevertheless, as demonstrated in Table 6, the highest average similarity observed between these two OC systems is 58.80% for the SBERT method. In addition, the SBERT method is the most effective, with a total of 229 matches.

**Table 7.** Worked Example (Scotland-England). Scores and validation labels

Original title	BoW	TF-IDF	SBERT	Ensemble	Mapping	Label
Staff Nurse Band 5	72%	68%	89%	82%	Registered Nurse	Correct
Senior Physiotherapist	45%	52%	91%	75%	Physiotherapist	Correct
Ward Manager	38%	41%	76%	58%	Service Manager	Partial
Cardiac Physiologist	22%	28%	67%	48%	Healthcare Science	Incorrect
Admin Officer Band 3	31%	35%	54%	43%	Admin Officer	Correct

At the job family level, OC system of the NHS in Northern Ireland has a four-character alphanumeric code followed by the job title. The length of the job titles is considerably greater than in the other three countries, and the number of unique job titles is also greater than in the ISCO and SOC systems. However, although the complex code's nomenclature, the SBERT method can identify 252 matches as shown in [Table 6](#).

### Standardization of job titles

Variable coding is essential in processing administrative registers. [Clausen \(2015\)](#) corrected the spelling of job titles in the Danish Demographic Database to create a list of all possible occupation variations to be coded in ISCO. While ISCO classifies job titles according to tasks and responsibilities, the Australian and New Zealand Standard Classification of Occupations (ANZSCO) characterizes jobs by skill level ([Eagers, Franklin, Yau, & Broome, 2018](#)). Standardization makes it easier to adopt SOC in the NHS.

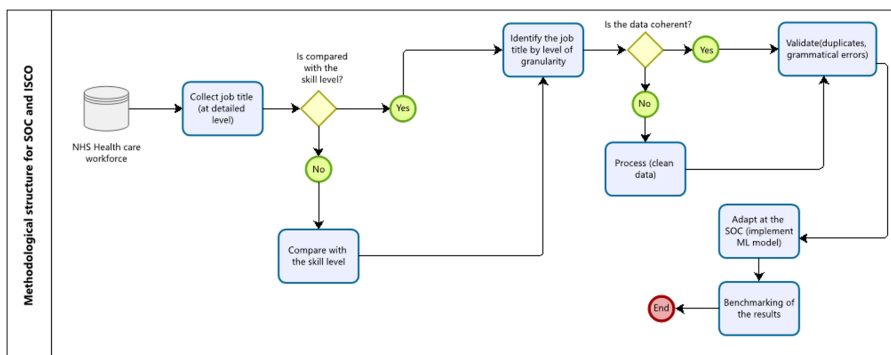
### Feasibility to adopt the SOC in the NHS

Benchmarking between the UK OC systems shows a significant level of heterogeneity between them. According to the [ONS \(2024\)](#), the SOC uses skill levels for its classification. Therefore, the NHS will need to adapt the job titles of staff currently working at a specific skill level.

To adapt the SOC in these organizations, a hierarchical approach will be required to classify staff job titles into hierarchical levels. One of the characteristics of the [ONS \(2024\)](#) SOC structure is its hierarchical division into four groups.

Based on the foregoing analysis, [Figure 3](#) proposes a pathway for adoption from the current NHS organizational classification system to achieve a high level of comparability with the SOC. The steps are: i) to collect job titles at a detailed level, ii) to compare with the skill level of education, iii) to create a hierarchical structure, and the location of every job title within a specific level, iv) to standardize the level, v) to apply a coherence analysis for cleaning the data, vi) to validate the data: duplicates, grammatical errors, and so on, vii) to implement SBERT model and viii) to benchmark the results (analyse the level of match by similarity range with the SOC).

This research establishes a foundational methodology for the adoption of SOC or ISCO-08 across the NHS in the United Kingdom and comparable healthcare systems around the world. Unlike CASCOT and OIM, where matching similar concepts is the basic framework, the Cascade methodology provides an ordered framework designed to address the high level of heterogeneity of the NHS ([Table 8](#)). However, it could be improved by creating a



**Figure 3.** Diagram of the process to adopt the NHS's organizations' OC system at the SOC. Beginning with the mapping of the NHS source to identify the hierarchical level, to make a cleaning and validation process, then to adapt at the SOC by NLP model, and finish with an analysis of the results by benchmarking technique

**Table 8.** Main differences among the methodology chosen “cascade method”, which was influenced by the heterogeneous OC systems in the United Kingdom, and the model SBERT regarding the level of granularity and the consistency of the results, at difference of CASCOT, where both systems SOC and ISCO have a similar level of granularity, and OIM as a merging of concepts

Cascade method	CASCOT	OIM
Hierarchical analysis, semantic mapping (NLP models) and benchmarking	Input SOC job titles, output ISCO codes. Based on a probability score (1–100), depending on the similarity algorithms	Mapping and Integration process (merging concepts). To sub-categorize in a significant way the existing ontological categories
Validation: Manual (100 samples/dataset, 2 annotators, Kappa 0.23–0.67)	Validation: Rule-based matching	Validation: Conceptual merging
Focus: UK healthcare sector, 4 NHS systems	Focus: General SOC-ISCO mapping	Focus: Generic ontology integration

comprehensive master list of NHS job titles and descriptions, and by involving a panel of professionals to examine and refine those mappings scoring below 30% similarity.

For practical deployment, we recommend: (1) automatic acceptance for matches >80% similarity with periodic spot-checking; (2) Expert review queue for matches 60–80% with domain specialist validation; (3) manual classification required for matches <60% or rejection below 30%. Model ownership should reside with NHS informatics teams with quarterly revalidation cycles. These thresholds directly support workforce decisions: >80% similarity enables cross-border staff deployment without additional verification; 60–80% informs training equivalence assessments; <60% triggers competency review before role recognition. Several risks require attention: (1) potential bias in occupational descriptions may affect model performance for underrepresented roles; (2) job title drift over time requires periodic retraining; (3) taxonomy evolution in both SOC and NHS systems necessitates ongoing maintenance; (4) this study demonstrates feasibility and a validation is recommended before deployment. For the Scotland–England comparison (100 job title mappings), 29 mappings (29%) exceeded the 80% similarity threshold and were auto-accepted, 33 (33%) fell within the 60–80% band and therefore required expert review and 38 (38%) scored below 60% and required manual classification. On the manually validated sample, the SBERT-based approach achieved an F1 score of 0.99, indicating near-perfect mapping performance. The 60–80% similarity band is of particular operational importance, as it identifies mappings that are not sufficiently robust for automation but can be resolved through expert review rather than full manual classification. Governance responsibility rests with NHS Digital, with annual threshold review and model retraining triggered when precision falls below 0.70. Collectively, this framework has the potential to reduce manual effort in routine workforce reconciliation while preserving oversight for lower-confidence mappings.

Future work is prioritized as follows: (1) Immediate: developing a real-time classification software application with expanded validation; (2) Near-term: integration with NHS workforce systems and automated model retraining; (3) Long-term: extension to the social care sector and international ISCO harmonization.

Implementation resources require minimal infrastructure: a single GPU server or cloud computing for model inference, plus 2–3 FTE workforce analysts for validation, with a timeline: 3-month pilot, 3-month validation, 6-month national rollout. Annual maintenance: 0.5 FTE for threshold recalibration and taxonomy updates when precision drops below 0.70.

## Conclusion

This study demonstrates the technical feasibility of harmonizing occupational classification systems within the UK healthcare sector using a three-stage cascade methodology involving

---

hierarchical analysis, semantic mapping with ensemble natural language processing (NLP) models and systematic benchmarking. Furthermore, this methodology enables a single standard to be used to achieve consistency in job title descriptions across the four UK systems.

Following [Krasna et al. \(2024\)](#) benchmarking approach, we identified matches at different varying confidence levels across the highest granularity tier. Specifically, UK pairwise comparisons used the following title bases: Scotland–Wales compared 84 non-medical and dental sub-job family titles against 82 third-level titles; Scotland–England compared 84 sub-job family titles against 61 fourth-level titles and Scotland–Northern Ireland compared 84 sub-job family titles against 626 job family titles. Although these title sets differ substantially in size and structure and are not directly normalized to a shared title base, the use of SBERT for semantic matching partially addresses this limitation by capturing meaning-level similarities between job titles regardless of differences in terminology, coding conventions or catalogue size across systems.

[UNECE \(2019\)](#) argues that common metadata and classifications are essential for cross-system comparability, yet each UK system maintains its own metadata, granularity and classification structure. Addressing this fragmentation requires future work towards a common business process for workforce data collection and training needs across the four nations. Such a framework should emphasize semantic interpretation within NHS organizations, for which SBERT embedding models have proven most suitable given their ability to capture meaning-level similarities across heterogeneous classification systems.

The practical value of this research lies in the provision of a replicable methodological framework and the demonstration of feasibility for an unaddressed challenge in UK health sector workforce analytics. While the contribution is incremental rather than algorithmically novel, it offers actionable guidance for practitioners. Critical next steps include formal validation with domain experts, the development of a governance framework, and the implementation of the programme in phases.

Here, we showed that using NLP models, particularly SBERT with validation metrics of 0.73–0.98 precision and 0.84–0.99 F1 scores, are useful tools for harmonizing OC in the UK health sector. Even though the Irish system’s nomenclature includes an alphanumeric code, the SBERT method can identify 252 matches with the Scottish system. In most cases, the SBERT model has the best performance, but there are cases with low similarity where the ensemble model can identify more matches. Both SBERT and ensemble models can be used to match similarities of job titles, but depending on the case, it could be more useful to use SBERT for matching job titles with a high similarity and an ensemble model for lower similarity. This research applies established NLP techniques to a health-sector context lacking prior methodological development.

This study can be applied in a scenario such as: COVID-19 redeployment required matching 2,847 Scottish staff to English equivalents. Our 89% similarity threshold enabled automated recognition of Band 5 Staff Nurse equivalence.

### **Acknowledgments**

The author would like to thank Dr Colin Tilley, Programme Director at NHS Education for Scotland (NES), for his invaluable contributions to the occupational health workforce in Scotland. The author would also like to thank Dr Diego Morales, Research Professor in Statistics and Probability, for his significant contributions and insightful suggestions during the development of this article. The author is grateful to the reviewers for their insightful comments and suggestions, which substantially strengthened the manuscript.

### **Supplementary material**

The supplementary material for this article can be found online

## References

- Abraham, J. D., Lambert, D. D., Mihalecz, M. C., Elcott, M. D., Asbury, H. S., & Palmer, P. C. (2023). How similar is similar enough? Job profile similarity benchmarks using occupational information network data. *International Journal of Selection and Assessment*, 31(3), 469–476. doi: [10.1111/ijsa.12430](https://doi.org/10.1111/ijsa.12430).
- Abubakar, H. D., Umar, M., & Bakale, A. (2022). Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec. *SLU Journal of Science and Technology*, 4(1), 27–33.
- Ahuir, V., Hurtado, L. F., García-Granada, F., & Sanchis, E. (2023). ELiRF-VRAIN at PoliticES-IberLEF2023: Dealing with long texts in transformer-based systems for user profiling. In *IberLEF@ SEPLN*.
- Cheong, C. W., Yin, K., Cheung, W. K., Fung, B. C., & Poon, J. (2023). Adaptive integration of categorical and multi-relational ontologies with EHR data for medical concept embedding. *ACM Transactions on Intelligent Systems and Technology*, 14(6), 1–20. doi: [10.1145/3625224](https://doi.org/10.1145/3625224).
- Cichosz, P. (2023). Bag of words and embedding text representation methods for medical article classification. *International Journal of Applied Mathematics and Computer Science*, 33(4), 603–621. doi: [10.34768/amcs-2023-0043](https://doi.org/10.34768/amcs-2023-0043).
- Clausen, N. F. (2015). The Danish demographic database—principles and methods for cleaning and standardisation of data. In *Population Reconstruction* (pp. 3–22). Cham: Springer International Publishing.
- Dai, S., Li, K., Luo, Z., Zhao, P., Hong, B., Zhu, A., & Liu, J. (2024). AI-based NLP section discusses the application and effect of bag-of-words models and TF-IDF in NLP tasks. *Journal of Artificial Intelligence General science (JAIGS)*, 5(1), 13–21. doi: [10.60087/jaigs.v5i1.149](https://doi.org/10.60087/jaigs.v5i1.149).
- Department of Health (2024). 'Northern Ireland health and social care (HSC) workforce statistics June 2023 to December 2023. Department of Health. Available from: <https://www.health-ni.gov.uk/articles/staff-numbers> [accessed 24 June 2024].
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies* (Vol. 1, pp. 4171–4186).
- Dhawaleswar Rao, C. H., & Pani, P. (2025). Advanced semantic text similarity analysis using sentence transformers. In *International Conference on Intelligent Computing and Communication* (Vol. 1240, pp. 371–379). doi: [10.1007/978-981-96-1264-2\\_33](https://doi.org/10.1007/978-981-96-1264-2_33).
- Eagers, J., Franklin, R. C., Yau, M. K., & Broome, K. (2018). Pre-retirement job and the work-to-retirement occupational transition process in Australia: A review. *Australian Occupational Therapy Journal*, 65(4), 314–328. doi: [10.1111/1440-1630.12452](https://doi.org/10.1111/1440-1630.12452).
- Fries, J. A., Steinberg, E., Khattar, S., Fleming, S. L., Posada, J., Callahan, A., & Shah, N. H. (2021). Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communications*, 12(1), 2017. doi: [10.1038/s41467-021-22328-4](https://doi.org/10.1038/s41467-021-22328-4).
- Ge, C., Friesen, M. C., Locke, S. J., Russ, D. E., Burstyn, I., Baker, C. J., & Huss, A. (2023). Automated coding of job descriptions from a general population study: Overview of existing tools, their application and comparison. *Annals of Work Exposures and Health*, 67(5), 663–672.
- Giunchiglia, F., & Shvaiko, P. (2003). Semantic matching. *The Knowledge Engineering Review*, 18(3), 265–280. doi: [10.1017/S0269888904000074](https://doi.org/10.1017/S0269888904000074).
- Graff, M., Moctezuma, D., & Téllez, E. S. (2025). 'Bag-of-Word approach is not dead: A performance analysis on a myriad of text classification challenges. *Natural Language Processing Journal*, 11, 100154. doi: [10.1016/j.nlp.2025.100154](https://doi.org/10.1016/j.nlp.2025.100154).
- Guleria, P., Frnda, J., & Naga Srinivasu, P. (2025). NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis. *Array*, 27, 100467. doi: [10.1016/j.array.2025.100467](https://doi.org/10.1016/j.array.2025.100467).

- 
- ILOSTAT-1 (2024). *International standard classification of occupations (ISCO)*. International Labour Organization. Available from: <https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation> [accessed 17 June 2024].
- Jalilifard, A., Caridá, V. F., Mansano, A. F., Cristo, R. S., & da Fonseca, F. P. C. (2021). Semantic sensitive TF-IDF to determine word relevance in documents. In Thampi, S. M., Gelenbe, E., Atiquzzaman, M., Chaudhary, V., & Li, K. C. (Eds), *Advances in Computing and Network Communications. Lecture Notes in Electrical Engineering*. Singapore: Springer. doi: [10.1007/978-981-33-6987-0\\_27](https://doi.org/10.1007/978-981-33-6987-0_27).
- Kouretsis, A., Bampouris, A., Morfiris, P., & Papageorgiou, K. (2020). labourR: classify multilingual labour market free-text to standardized hierarchical occupations. *R package*.
- Krasna, H., Venkataraman, M., Robins, M., Patino, I., & Leider, J. P. (2024). Standard occupational classification codes: Gaps in federal data on the public health workforce. *American Journal of Public Health, 114*(1), 48–56. doi: [10.2105/AJPH.2023.307463](https://doi.org/10.2105/AJPH.2023.307463).
- Kuodytė, V., & Petkevičius, L. (2021). Education-to-skill mapping using hierarchical classification and transformer neural network. *Applied Sciences, 11*(13), 5868. doi: [10.3390/app11135868](https://doi.org/10.3390/app11135868).
- Li, J., Qi 'na, F. and Kuo, Z. (2007, In this issue). Keyword extraction based on tf/idf for Chinese news document. *Wuhan University Journal of Natural Sciences, 12*(5), 917–921. doi:[10.1007/s11859-007-0038-4](https://doi.org/10.1007/s11859-007-0038-4).
- Lwin, N., Jain, R., Dal, R., Yan, H., Thaw, K., & Naung, S. (2025). Text classification for Clickbait detection: A model-driven approach using CountVectorizer and ML classifiers. *Journal of Applied Science and Technology Trends, 06*(01), 43–49. doi: [10.38094/jastt61237](https://doi.org/10.38094/jastt61237).
- Mannetje, A. T., & Kromhout, H. (2003). The use of occupation and industry classifications in general population studies. *International Journal of Epidemiology, 32*(3), 419–428. doi: [10.1093/ije/dyg080](https://doi.org/10.1093/ije/dyg080).
- NHS England (2024). NHS Workforce Statistics - December 2023, Available from: <https://digital.nhs.uk/data-and-information/publications/statistical/nhs-workforce-statistics> [accessed 22 June 2024].
- ONS (2024). The current standard occupational classification for the UK, published in three volumes, Available from: <https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020> [accessed 19 June 2024].
- Prasad, A., Mohammad Alenazy, W., Ahmad, N., Ali, G., Abdallah, H., & Ahmad, S. (2025). Optimizing IoT intrusion detection with cosine similarity based dataset balancing and hybrid deep learning, *15*(1), 30939. doi: [10.1038/s41598-025-15631-3](https://doi.org/10.1038/s41598-025-15631-3).
- Python Software Foundation (2023). *Python 3.11.4*. Anaconda. Available from: <https://www.python.org/> [accessed 14 October 2025].
- Rana, V., & Singh, G. (2014). An analysis of semantic heterogeneity issues and their countermeasures prevailing in semantic web. In *International Conference on Reliability Optimization and Information Technology (ICROIT)*, Faridabad (pp. 80–85). IEEE. doi: [10.1109/ICROIT.2014.6798296](https://doi.org/10.1109/ICROIT.2014.6798296).
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Available from: <https://arxiv.org/abs/1908.10084>
- Russ, D. E., Ho, K.-Y., Colt, J. S., Armenti, K. R., Baris, D., Chow, W. H., . . . & Friesen, M. C. (2016). Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. *Occupational and Environmental Medicine, 73*(6), 417–424. doi: [10.1136/oemed-2015-103152](https://doi.org/10.1136/oemed-2015-103152).
- Russ, D.E., Josse, P., Remen, T., Hofmann, J.N., Purdue, M.P., Siemiatycki, J., Silverman, D.T., Zhang, Y., Lavoué, J., & Friesen, M.C. (2023). Evaluation of the updated SOCcer v2 algorithm for coding free-text job descriptions in three epidemiologic studies. *Annals of Work Exposures and Health, 67*(6), 772–783. doi:[10.1093/annweh/wxad020](https://doi.org/10.1093/annweh/wxad020).

- Safikhani, P., Avetisyan, H., Föste-Eggers, D., & Broneske, D. (2023). Automated occupation coding with hierarchical features: A data-centric approach to classification with pre-trained language models. *Discover Artificial Intelligence*, 3(1), 6. doi: [10.1007/s44163-023-00050-y](https://doi.org/10.1007/s44163-023-00050-y).
- Santander-Cruz, Y., Salazar-Colores, S., Paredes-García, W. J., Guendulain-Arenas, H., & Tovar-Arriaga, S. (2022). Semantic feature extraction using SBERT for Dementia detection. *Brain Sciences*, 12(2), 270. doi: [10.3390/brainsci12020270](https://doi.org/10.3390/brainsci12020270).
- Schmidtke, H. R. (2014). Context and granularity. In *Context in Computing: A Cross-Disciplinary Approach for Modeling the Real World* (pp. 455–470). New York: Springer. doi: [10.1007/978-1-4939-1887-4\\_28](https://doi.org/10.1007/978-1-4939-1887-4_28).
- Shahzad, S. K., Ahmed, D., Naqvi, M. R., Mushtaq, M. T., Iqbal, M. W., & Munir, F. (2021). Ontology driven smart health service integration. *Computer Methods and Programs in Biomedicine*, 207, 106146. doi: [10.1016/j.cmpb.2021.106146](https://doi.org/10.1016/j.cmpb.2021.106146).
- StatsWales (2024). Important update for StatsWales OData users, NHS staff, Available from: <https://statswales.gov.wales/Catalogue/Health-and-Social-Care/NHS-Staff> [accessed 24 June 2024].
- Suarez Garcia, C. A., Adishes, A., & Baker, C. J. O. (2021) (In this issue). S-464 Automated occupational encoding to the Canadian national occupation classification using an ensemble classifier from TF-IDF and Doc2Vec embeddings. *Occupational and Environmental Medicine*, 78(Suppl. 1), A161–A161. doi: [10.1136/OEM-2021-EPI.442](https://doi.org/10.1136/OEM-2021-EPI.442).
- Susan, S., Sharma, M., & Choudhary, G. (2024). Uniqueness meets semantics: A novel semantically meaningful bag-of-words approach for matching resumes to job profiles. *Journal Iberamia*, 24 (74), 117–132. doi: [10.4114/intartif.vol24iss74](https://doi.org/10.4114/intartif.vol24iss74).
- Tijdens, K. G., & Kaandorp, C. S. (2018). Validating occupational coding indexes for use in multi-country surveys. In *Survey Methods: Insights from the Field* (pp. 1–12). doi: [10.13094/SMIF-2018-00007](https://doi.org/10.13094/SMIF-2018-00007).
- Turas Data Intelligence (2024). NHS Scotland workforce (phase one), Available from: <https://turasdata.nes.nhs.scot/data-and-reports/official-workforce-statistics/all-official-statistics-publications/04-june-2024-workforce/dashboards/nhs-scotland-workforce-phase-one/?pageid=11804> [accessed 21 June 2024].
- UNECE (2019). Generic statistical Business Process Model. Version 5.1 January 2019, Available from: <https://unece.org/statistics/documents/2019/01/standards/gsbpm-v51> [accessed 28 June 2024].
- Warwick (2018). Cascot: Computer assisted structured coding tool, Available from: [https://warwick.ac.uk/fac/soc/ier/data\\_group/cascot](https://warwick.ac.uk/fac/soc/ier/data_group/cascot) [accessed 20 June 2024].
- Yilmaz, Y., Jurado Nunez, A., Ariaeinejad, A., Lee, M., Sherbino, J., & Chan, T. M. (2022). Harnessing natural language processing to support decisions around workplace-based assessment: Machine learning study of competency-based medical education. *JMIR Med Educ*, 8(2), e30537. doi: [10.2196/30537](https://doi.org/10.2196/30537). Available from: <https://mededu.jmir.org/2022/2/e30537>
- Young, T. K., Fedkina, N., Chatwood, S., & Bjerregaard, P. (2018). Comparing healthcare workforce in circumpolar regions: Patterns, trends and challenges. *International Journal of Circumpolar Health*, 77(1), 1492825. doi: [10.1080/22423982.2018.1492825](https://doi.org/10.1080/22423982.2018.1492825).
- Zhang, L. Y., Ren, J. D., & Li, X. W. (2017). OIM-SM: A method for ontology integration based on semantic mapping. *Journal of Intelligent and Fuzzy Systems*, 32(3), 1983–1995. doi: [10.3233/jifs-161553](https://doi.org/10.3233/jifs-161553).

### Corresponding author

Sonia Elizabeth Herrera-Salgado can be contacted at: [elizabetharcangel@hotmail.com](mailto:elizabetharcangel@hotmail.com)