3 Online job postings as a data source to analyse the impact of digitalisation on labour markets

This chapter discusses how big data and the information contained in millions of job postings collected from the internet can be used to track the evolution of labour markets and skill demands. The chapter discusses the advantages and the disadvantages of using big data for labour market intelligence over traditional statistics. The chapter also provide a description of the natural language processing approach used to analyse the information contained in online job postings and the metrics that are produced and presented throughout the report.

Every day, millions of individuals around the globe use new technologies to search for a job. Web platforms such as LinkedIn, Monster, Indeed, ZipRecruiter or CareerBuilder aggregate the information of millions of users and firms who meet daily in this marketplace. All those platforms provide their users with an "electronic labour market" where millions of new jobs are advertised every day.

New advancements in automated web scraping technologies (i.e. the automated retrieval and storage of textual information from the internet) allow to collect the information contained in job postings that have been published online and use it to analyse trends in labour market dynamics and skill demands. The advantages of using the information contained in online job postings over traditional labour market statistics lie in its richness, timeliness and granularity (OECD, 2021[1]).

First of all, the information contained in online job postings is collected on a rolling basis, allowing to track the evolution of skills demanded by employers up to very recent months and to detect new and emerging trends as well as technologies that may be growing and in high-demand. In particular, unlike other data sources that are based on the collection of survey information that is updated only with significant lags (i.e. O*NET or ESCO), the analysis of online job postings allows to track the changes in skill demands over time, up to very recent months. The study of online job postings also allows to examine the cross-sectional variation in skill requirements within occupations where skill demands for the same occupation may vary depending on the geography analysed. Both, in turn, allow to capture the changing nature of skill demands and the heterogeneous impact of technological progress across occupations.

Second, the granularity and high volume of the information contained in online vacancies allows to move from the analysis of generic concepts such as the assessment of the demand for the "Knowledge of Informatics" (assessed in other databases like O*NET) to the estimation of the impact on labour market of much more granular and specific knowledge domains such as "Python programming" or "Web design".

This report makes use of a large database of job postings published online and collected by Lightcast (n.d.[2]). For instance, Lightcast mines and codes millions of job postings from more than 40 000 online sources daily, scraping up to 3.4 million active job postings from thousands of webpages in the United States. Similarly, in European countries, Lightcast has more than 900 scrapers/robots monitoring more than 35 000 job portals every day and collecting more than 1 million new job postings daily.

The information contained in the Lightcast database provides up to 70 different variables ranging from skill keywords contained in each job posting, qualifications and experience required to fill the job and its geographical location, the name of the firm that is advertising the vacancy as well as the type of contract (permanent, temporary) and, when available, the salary offered for the specific role advertised.

The data are presented by a unique job-identifier and the deduplication of job postings appearing in different web and career portals ensures that the same job is not counted more than once even if appearing in different web-portals. Job postings are then mapped to different taxonomies and, in particular, to the Standard Classification of Occupations (SOC) at the 6 digits disaggregation level.

Lightcast also puts considerable effort in harmonising the skills found in the job postings. For instance, skill keywords such as "teamwork" and "collaboration" are combined into "teamwork/collaboration", and words that have several accepted spellings are considered interchangeably. It is important to notice that not all keywords collected from job postings are "skills" *strictu sensu*. Many represent "knowledge areas" (i.e. Endocrinology or Mathematical Modelling), others identify the use of specific "technologies and tools" (i.e. Python or Excel) while others relate to "abilities" required to perform an occupation (i.e. Physical Abilities or Cognitive Abilities). While the distinction between these categories bears meaningful information, this study pools them together in the analysis and distinguishes between the different concepts only when appropriate. For the sake of simplicity, in the remainder of this study, the term "skills" will be used when referring to all these different dimensions globally while knowledge, abilities, technologies and tools will be used in italics to clearly distinguish between the different concepts when necessary.

Box 3.1. Knowledge, skills, abilities, technologies and tools: What is what?

Knowledge keywords refer to an organised body of information usually of a factual or procedural nature which, if applied, makes adequate performance on the job possible. Examples are keywords such as Endocrinology which, in job postings, denote the required knowledge of all different aspects related to the medical discipline related to it and to the body of information that relates to it directly.

Skill keywords refer to the proficient manual, verbal or mental manipulation of data or things. Skills can be readily measured by a performance test where quantity and quality of performance are evaluated, usually within an established time limit. Examples of proficient manipulation of things are skill in typing or skill in operating a vehicle. Examples of proficient manipulation of data are skill in computation using decimals; skill in editing for transposed numbers, etc.

Ability keywords refer to the power to perform an observable activity at the present time. This means that abilities have been evidenced through activities or behaviours that are similar to those required on the job (e.g. ability to plan and organise work).

Technology and tool keywords refer to the knowledge of and ability to utilise certain technologies in a work context. Keywords such as Python, for instance, refer to the required knowledge of that software programming language which can be applied to tasks in different occupations. Similarly, keywords such as Excel, refer to the ability of using that statistical software package in a work-setting.

In the remainder of this report, the word "skills" is used to refer to all the above dimensions, unless more precision is needed and the specific terms are, therefore, used instead.

Source: OECD (2017[3]), "*Getting Skills Right: Skills for Jobs Indicators,*" https://dx.doi.org/10.1787/9789264277878-en.

Online job postings are also mapped to different national and international taxonomies. Information for the United States is, for instance, mapped to the US Standard Occupational Classification (SOC). In the United Kingdom, information is mapped to the UK SOC taxonomy while the National Occupational Classification (NOC) is used in Canada. On top of these country-specific occupational taxonomies, in Anglophone countries analysed in this report (namely Canada, Singapore, UK and US), Lightcast provides an overarching proprietary occupational taxonomy that allows the easy comparison of statistics computed at the occupation level across those countries.

The International Standard Classification of Occupations (ISCO) is, instead, used to classify European countries' job postings. Lightcast, however, does not provide a cross-walk between its proprietary occupational taxonomy (for Anglophone countries) and ISCO (for EU countries) so that this report, when analysing statistics at the occupation level, will refer to both the Lightcast taxonomy as well as to ISCO when applicable.

The use of different occupational taxonomies can create some challenges when comparing statistics across Anglophone countries[1] (using Lightcast occupational taxonomy) and European countries (using the ISCO taxonomy).[2] This report tries to minimise these issues by selecting occupations that are relatively comparable to each other across taxonomies. For instance, Chapter 3 analyses "database administrators" in Anglophone countries using Lightcast taxonomy at the 8th digit disaggregation level and "database designers and administrators" in EU countries using the ISCO taxonomy at the 4th digit disaggregation level. The two occupations are logically comparable, but the statistics for the latter are presented at a higher level of aggregation, meaning that more job roles may be contained therein than in the corresponding case for countries using Lightcast taxonomy.

This is also to say that the highest level of granularity in the Lightcast occupational taxonomy allows to go deeper into the analysis of smaller-sized and more detailed occupations relative to ISCO. This certainly enriches the analysis but it also calls for caution when comparing the statistics of some of the smaller-sampled occupations (in Lightcast taxonomy) with those of occupational groups that encompass more job roles (in ISCO) and that may pool together a larger number of job postings.

While the wealth and granularity of skills and labour market information contained in online vacancies is unprecedented, caveats and limitations to the use of this data also exist. First, not all job adverts are actually published online and therefore statistics therein may not be representative of the whole labour market and of "off-line" job openings. As pointed out by (Hershbein and Kahn, 2016[4]) vacancies appearing online are likely to be skewed towards certain areas of the economy despite the fact that available jobs have been increasingly appearing online instead of in traditional sources, such as newspapers. On these regards, (Carnevale, Jayasundera and Repnikov, 2014[5]) estimate that around 80-90% of postings requiring at least a Bachelor's degree can be found online, whereas 40-60% of job postings requiring a high school degree are channelled through the internet. That being said, (Hershbein and Kahn, 2016[4]) also suggest that when comparing the relative frequency of postings in online vacancies data to survey-based data online vacancies, online vacancy data reflect labour demand reasonably well and that the differences that emerge appear relatively stable over time. Recent OECD studies also highlight that the potential bias is likely to be more pronounced in low skilled jobs and less of a concern for high-skilled occupations and sectors (Cammeraat and Squicciarini, 2021[6]) like the one analysed in this report. That being said, not all high skilled vacancies are posted online and some are channelled through informal labour market networks (see Box 3.2)

Online vacancies may also tend to under-report certain skills as they may be "implicit" in job postings but still equally important for carrying out the tasks of the occupation at hand. This can produce an upward bias in the frequency with which certain cognitive, technical or soft skills appear relative to physical or routine skills that are less frequently mentioned.

Box 3.2. The role of informal hiring networks

The role of "informal hiring networks" can remain unaccounted for when analysing online job postings. Qualitative research done by Randstad Research Italy (forthcoming[7]) shows that these networks can be key in the case of backend programming profile (from cybersecurity to software coding). This leads to think that such roles and professions might be growing at an even faster rate than online job postings' analysis suggests in a context where demand is particularly high and job retention low. Randstad's qualitative study shows that companies looking to hire such roles need strong hiring channels, alternative to online job posting, such as educational/personal networks. In the case of start-ups for example, OJPs and HR services are utilised for the hiring of a first professional figure, usually in a senior role, who subsequently utilises their network to directly attract suitable candidates.

It is also important to notice that not all job postings collected by Lightcast contain all fields of information. Wages are, for instance, available only for a subset of vacancies.

This report and the data used therein cover 10 countries: Canada, Belgium, France, Germany, Italy, the Netherlands, Spain the United Kingdom and the United States.

Time series for Canada, Singapore, the United Kingdom and the United States is available starting from the year 2012 while information on online job postings for Belgium, France, Germany, Italy and the Netherlands starts in 2018 up to recent months of 2022. This report analyses the evolution of job postings also during the recent COVID-19 pandemic. Results, therefore, capture the significant drop in economic activity and the associated labour market disruptions starting in early 2020 when virtually all countries put in place severe measures to control the spread of the Sars-Cov-19 virus.

3.1. Using machine learning to analyse the information contained in online job postings

The information contained in online job postings is extremely rich and large. The database used in this paper spans several gigabytes of data and sums up to millions of keywords collected from job postings in different countries and over time. In addition to its size, the information contained in online job postings differs from most traditional labour market statistics (such as, for instance, labour force surveys) in that it contains information in the form of text rather than numbers and figures. Differently from standard quantitative data, text bears "semantic meaning" which can be multifaceted and ambiguous but it can also convey a far greater amount of information than just numbers and figures.

Recent advances in machine learning techniques led to the development of so-called language models which have the objective of understanding the complex relationships between words (their semantics) by deriving and interpreting the context those words appear in. Language models (and in particular Natural Language Processing- NLP- models) interpret text information by feeding it to machine learning algorithms that derive the logical rules to interpret the semantic context in which words appear. NLP and language models, used in the remainder of this paper, are therefore better suited for the analysis of text information than traditional statistics and, as such, they are used for the analysis of online job postings in the remainder of this report.

In particular, the approach taken in this report leverages "Word2Vec", a NLP algorithm developed by researchers in Google (Mikolov et al., 2013[8]). This algorithm functions by creating a mapping between the meaning (i.e. the semantics) of words contained in text and mathematical vectors, so-called "word vectors". Put it differently, word vectors are the mathematical representation of the meaning of the words used in online job postings. Those vectors are plotted in a high-dimensional vector space (called "graph") where words with similar meanings occupy close spatial positions in the vector space (see Annex 3.A).

Since word vectors[3] occupy a specific place in the vector space, this makes it possible to calculate the distance (i.e. the cosine similarity) between those vectors and to rank the relationships between skills from the closest to the farthest from any given occupation. In other words, by estimating their semantic closeness, this approach allows to rank the similarities between every skill (word) vector relative to any given occupation vector.[4]

Skills that are more similar to a certain occupation are interpreted in this report as being more "relevant" to the occupation (see Annex 3.A). Using this approach is, therefore, possible to assess whether the skill "Excel" is more relevant to the occupation "Economist" or to "Painter", based on the semantic closeness of these words' meanings extrapolated from millions of job postings.

In the remainder of this report, the matrix of skills-to-occupations relevance scores (the Semantic Skill Bundle Matrix, SSBM) is used to identify the occupations for which digital skills are particularly relevant as well as to assess the relationship between digital skills and occupations and the speed of diffusion of the demand for digital technologies and skills across labour markets. Technical details about the machine learning approach used are provided in Annex 3.A.

It is important to notice that this report focuses on a selection of digital occupations, skills and technologies. Those have been identified by OECD experts and researchers in the Randstad Research Institute. The list of digital occupations and skills has been drafted trying to strike a balance between the need to analyse the impact of the digitalisation in various parts of the labour markets and the need to have a focused and targeted analysis of the phenomena at hand. Next chapter lists and describes the occupations and skills analysed in this report and reasoning behind their selection for the report.

References

Boleda, G. (2020), "Distributional Semantics and Linguistic Theory", *Annual Review of Linguistics*, Vol. 6/1, pp. 213-234, https://doi.org/10.1146/annurev-linguistics-011619-030303. [11]

Cammeraat, E. and M. Squicciarini (2021), "Burning Glass Technologies' data use in policy-relevant analysis: An occupation-level assessment", *OECD Science, Technology and Industry Working Papers*, No. 2021/05, OECD Publishing, Paris, https://doi.org/10.1787/cd75c3e7-en. [6]

Carnevale, A., T. Jayasundera and D. Repnikov (2014), *The online college labor market: Where the jobs are*, Georgetown University, https://cew.georgetown.edu/wp-content/uploads/2014/11/OCLM.Exec_.Web_.pdf. [5]

Erk, K. (2012), "Vector Space Models of Word Meaning and Phrase Meaning: A Survey", *Language and Linguistics Compass*, Vol. 6/10, pp. 635-653, https://doi.org/10.1002/lnco.362. [10]

Harris, Z. (1954), "Distributional Structure", *WORD*, Vol. 10/2-3, pp. 146-162, https://doi.org/10.1080/00437956.1954.11659520. [9]

Hershbein, B. and L. Kahn (2016), *Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings*, National Bureau of Economic Research, Cambridge, MA, https://doi.org/10.3386/w22762. [4]

Jurafsky, D. and J. Martin (2021), *Speech and Language Processing*, https://web.stanford.edu/~jurafsky/slp3/. [12]

Lightcast (n.d.), *Lightcast website*, https://lightcast.io/. [2]

Mikolov, T. et al. (2013), *Efficient Estimation of Word Representations in Vector Space*. [8]

OECD (2021), *OECD Skills Outlook 2021: Learning for Life*, OECD Publishing, Paris, https://doi.org/10.1787/0ae365b4-en. [1]

OECD (2017), *Getting Skills Right: Skills for Jobs Indicators*, Getting Skills Right, OECD Publishing, Paris, https://doi.org/10.1787/9789264277878-en. [3]

Randstad Research Italy (forthcoming), *Connessioni al servizio della fruibilità. Le 100 e più professioni digitali del futuro*. [7]

Annex 3.A. Further insights on the machine learning approach to analyse to extract skill relevance scores

Previous literature that used online vacancies to analyse labour market dynamics has, in most cases, done so via frequency-based measures and by counting the number of times certain skills would be mentioned in online job postings. Recent developments in Natural Language Processing (NLP), however, allow to leverage the information contained in online vacancies in a much more sophisticated way by looking at the semantic meaning of the textual information contained in online job postings. One such approach, the so-called word embeddings, derive a word's meaning from the context this occurs in.

These sophisticated methods leverage the distributional hypothesis, as stated by (Harris, 1954[9]), use the context in which a certain term occurs to derive the semantic meaning of the term. In their most common form, vector space models use the word's context to derive the meaning of a word and create *n*-dimensional vectors to represent that meaning. This is a so-called semantic representation which is thus encoded and distributed over all the *n* dimensions of the vector, where each dimension stands for a certain context item and its co-ordinates refer to the count of this context (Erk, 2012[10]). This quantification of the semantics of words allows to compute mathematical similarity measures that reflect the similarity between different vectors representing different words and concepts (Boleda, 2020[11]).

Intuitively word vectors are retaining the semantic meaning of words and algebraic operations can be performed using them. As word vectors retain the semantic meaning of their underlying words, the results of such mathematical operations are expected to also return semantically and logically meaningful results. For instance, once word vectors have been estimated, one could perform basic arithmetic using those vectors, such as:

$$vec(\text{“Chief”}) + vec(\text{“Male”}) + vec(\text{“Royalty”}) \approx vec(\text{“King”})$$

From a mathematical point of view, this means that if two words share an interrelated meaning (for example Chief, Male, Royalty and King) the cosine of the angle between their vector representations should be close to 1, i.e. the angle close to 0. Furthermore, negative values for the cosine refer to vector representations similar, but opposite in meaning such that vec(“King”)- vec(“Male”) ≈vec(“Queen”).

Based on these properties, one can compute measures of semantic similarity of pairs of skill keywords and occupations using paragraph vector distributed bag of words (PV-DBOW) to create the vectors representing occupations instead of simple skill keywords.[5]

Next, one can construct a Semantic Skill Bundle Matrix (SSBM) by calculating the similarity between all possible combinations of skills keywords and any given occupation pairs. Comparison of the vectors is done by looking at the so-called similarity and/ or distance between vectors. This is to say that, given two vectors representations for keyword A, and occupation B, the calculation of the similarity is done as:

$$\text{distance}(A, B) = (A \cdot B)/||A||\ ||B||$$

To illustrate the type of information contained in the SSBM, an example is given in Annex Table 3.A.1 for two randomly selected occupations “Web-Designer” and “Marketing Manager”.

Annex Table 3.A.1. Example of SSBM values for the occupation Web Designer and Marketing Manager

Web Designer (1)		Marketing Manager (2)	
Web Design	0.73	Online Marketing	0.57
Bootstrap	0.62	Marketing Management	0.52
Graphic And Visual Design	0.55	General Marketing	0.52
User Interface And User Experience	0.55	Marketing Strategy	0.50
Digital Design	0.55	Web Analytics	0.49
Javascript And Jquery	0.55	Media Strategy And Planning	0.47
Animation And Game Design	0.53	Content Development And Management	0.45
…		….	
Electrical Engineering Industry	-0.06	Civil Aviation Authority	-0.04
Occupational Hygiene	-0.06	Fuel metres	-0.04
Oil Well Intervention	-0.06	Diagnostic Technologies	-0.04
Oil Wells	-0.06	Repair	-0.06
Mechanical Products Industry Knowledge	-0.08	Thermoplastic	-0.07
Health Care Industry Knowledge	-0.11	Radio Frequency Equipment	-0.08

Note: Values reported in the table represent the cosine similarity between each skill keyword vector representation listed in column (1) and (2) and the vector representation of the occupation web designer and marketing manager. Higher values of the cosine similarity reflect higher semantic relatedness and it is interpreted as an indication of the relevance of the skill keyword for the occupation at hand.
Source: OECD calculations based on Lightcast data for the United Kingdom in 2018.

From an intuitive point of view, the closer (semantically, in meaning) a skill keyword vector representation is to the vector representation of the occupation and the more the skill is assumed to be relevant for the occupation at hand. Results in Annex Table 3.A.1 show that the skill vectors “Web Design”, “Bootstrap” and “Graphic and Visual Design” are semantically close to the occupation “Web Designer” and, hence, are interpreted in this paper as “relevant” to that occupation. Similarly, “Online Marketing”, “Marketing Management” and “General Marketing” are the most relevant skill keywords for “Marketing Managers”.

As noted by (2021[12]), intrinsic evaluation of word embeddings as the one discussed above is commonly evaluated by correlating the similarity scores with expert constructed scores. This annex provides empirical support to the hypothesis for which these SSBM scores can indeed be used to represent the relevance of skills for occupations by comparing them with the skill information contained in the O*NET database.

O*NET is a tool for career exploration and job analysis which contains detailed descriptions of more than 900 occupations and their corresponding knowledge, skills, abilities, and competencies. O*NET collects this data from job incumbents and occupation experts who assess the importance and level of each knowledge, skill, ability, and competency for each specific occupation and rank them correspondingly. Comparison between O*NET and SSBM data is useful in evaluating how well the SSBM represents the relations between occupations and skills and the cosine similarity scores represent skill relevance.

Direct comparison between SSBM scores and the O*NET values is, however, difficult to perform as the keywords in the SSBM (extracted directly from job postings) are far more granular and detailed than the categories analysed in the O*NET. For instance, in the SSBM skills such as “anaesthesiology” or “python” occur, while in the O*NET closest categories with underlying quantitative scores attached to them would be the far more aggregated knowledge areas of “Medicine and Dentistry” and “Computer and Electronics”.

Deriving a global measure for the alignment between the many SSBM dimensions and O*NET is, therefore, not a straightforward tasks. First, given the higher granularity of the keywords in the SSBM relative to O*NET, it is necessary to aggregate the data in the SSBM at the same level at that presented in O*NET.

This means, for instance, to group SSBM keywords such as Anaesthesiology, Patient care etc. into one group that can be compared with Medicine and Dentistry in O*NET.

The correlations across occupations between cosine similarity scores in the SSBM and O*NET ranked values can be used to aggregate the relevance values in the SSBM into O*NET categories. More specifically, this is done by weighting the SSBM relevance scores by the correlation of each skill keyword with every O*NET category using the values in and then taking the sum across all SSBM skills by occupation (see an example in Annex Box 3.A.1).

Annex Box 3.A.1. Mapping keywords from online job postings to O*NET categories

To describe the strategy used to carry out such grouping, let us assume, for simplicity that there are only two skills in the SSBM ("Aerospace Engineering", "Behaviour Analysis") and only two categories in O*NET, ("Education and Training" and "Engineering and Technology") one would first calculate the correlations between the SSBM and O*NET categories as follows:

Annex Table 3.A.2. Correlations between SSBM relevance scores and O*NET values

Correlations matrix between SSBM and O*NET values	Aerospace Engineering	Behavioural Analysis
Education and Training	-0.074	0.323
Engineering and Technology	0.515	-0.225

The simplified example in Annex Table 3.A.2 above indicates that occupations in the SSBM for which "Aerospace Engineering" is very relevant are also the same occupations (on average) where the knowledge of "Engineering and Technology" is also highly relevant in the O*NET (cross-occupation correlation= 0.51). On the contrary, occupations in the SSBM for which "Behavioural Analysis" is very relevant do not correlate with those for which "Engineering and Technology" in O*NET is highly relevant as one would expect.

The correlation information above can be used to aggregate the relevance values in the SSBM into O*NET categories by occupation. This is done by weighting the SSBM relevance scores by the correlation of each skill with every O*NET category using the values in Annex Table 3.A.2 and then taking the weighted sum across all SSBM skills by occupation.

In other words, consistent with the simplified example above, one can consider the occupation Engineering Technicians (SOC 17-3020) where the SBBM relevance scores for Engineering Technicians of the skills "Aerospace Engineering" and "Behaviour Analysis" are 0.570 and 0.092 respectively.

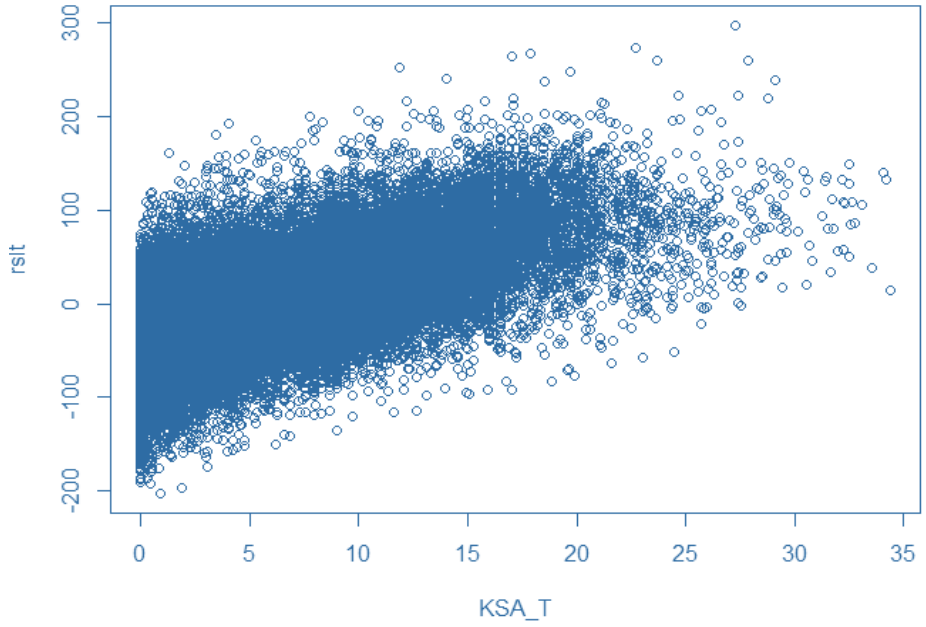
In order to translate these relevance scores into the O*NET categories of Education and Training and Engineering and Technology, one can take the sum of the SSBM relevance scores weighted by their correlation with the O*NET relevance values.

Once all skills in the SSBM have been grouped into the O*NET categories, it is possible to calculate the global correlation between SSBM (expressed in O*NET terms) and the O*NET ranked values.

In order to do so, an additional step is needed to calculate a unique measure of relevance in O*NET that combines "importance" and "level" to compare it with the relevance score in the SSBM across occupations. One can get this relevance score by multiplying the O*NET importance and level scores for each O*NET category across all occupations (OECD, 2017[3]).

Finally, this allows to correlate the O*NET values and the respective SSBM values across all combinations of occupations and skills. Annex Figure 3.A.1 shows that the correlation between the two variables is strong (0.62), positive and statistically significant providing further evidence of the alignment between the SSBM and O*NET values and of the validity of using SSBM relevance scores as an approximation of skill relevance for occupations.

Annex Figure 3.A.1. Global correlation between SSBM relevance scores and O*NET ranked values



Note: Dots represent occupations at the 6th digit US SOC level. Each dot is the combination of two values: on the horizontal axis (KSA_T) representing the O*NET scores (importance*level); on the vertical axis (rslt) the corresponding SSBM cosine similarity for every occupation.
Source: OECD calculations based on Lightcast data for the United States for the year 2019.

Notes

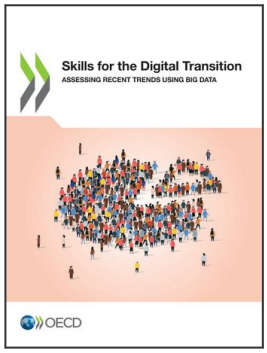
[1] The Anglophone countries covered in this report are: Canada, United Kingdom, United States and Singapore.

[2] The European countries covered in this report are: Belgium, France, Italy, Spain, the Netherlands

[3] One *n*-dimensional vector per skill.

[4] Occupation vectors are also calculated using a slight modification of Word2Vec called Doc2Vec (see Annex 3.A).

[5] PV-DBOW is an established technique in the natural language processing literature. A study by Mikolov et al. (2013) showed that in comparison to vector averaging (10.25%), bag-of-words (8.10%), bag-of-bigrams (7.28%) and weighted bag-of-bigrams (5.67%) the paragraph vectors, such as PV-DBOW, had the lowest error rate (3.82%). In contrast to other paragraph vector models, PV-DBOW ignores the context words in the input, but forces the model to predict words randomly sampled from the paragraph in the output. In short, each iteration of training, the model consecutively samples a text window and then a random word from this sampled text window to form a classification task for each given paragraph vector. Mikolov et al. (2013) state that in addition to being conceptually simple, PV-DBOW also requires less data to store than the distributed memory model (PV-DM).



From:

Skills for the Digital Transition

Assessing Recent Trends Using Big Data

Access the complete publication at:
https://doi.org/10.1787/38c36777-en

Please cite this chapter as:

OECD (2022), "Online job postings as a data source to analyse the impact of digitalisation on labour markets", in *Skills for the Digital Transition: Assessing Recent Trends Using Big Data*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/7d99dfbe-en