Organisation for Economic Co-operation and Development

DELSA/ELSA/WD/SEM(2021)10

Unclassified

English text only

9 November 2021

DIRECTORATE FOR EMPLOYMENT, LABOUR AND SOCIAL AFFAIRS
EMPLOYMENT, LABOUR AND SOCIAL AFFAIRS COMMITTEE

Speaking the Same Language: A Machine Learning Approach to Classify Skills in Burning Glass Technologies Data

JEL Codes: C45, C55, J23, J24, J63

Authorised for publication by Stefano Scarpetta, Director, Directorate for Employment, Labour and Social Affairs.

All Social, Employment and Migration Working Papers are now available through the OECD website at www.oecd.org/els/workingpapers.

Julie Lassébie: julie.lassebie@oecd.org
Luca Marcolin: luca.marcolin@oecd.org
Marieke Vandeweyer: marieke.vandeweyer@oecd.org
Benjamin Vignal: benjamin.vignal@ensae.fr

JT03485016

OECD Social, Employment and Migration Working Papers

http://www.oecd.org/els/workingpapers

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the author(s).

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to els.contact@oecd.org .

This series is designed to make available to a wider readership selected labour market, social policy and migration studies prepared for use within the OECD. Authorship is usually collective, but principal writers are named. The papers are generally available only in their original language – English or French – with a summary in the other.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Acknowledgements

This paper was prepared by the authors under the leadership of Stefano Scarpetta (Director, OECD Directorate for Employment, Labour, and Social Affairs – ELS), Mark Keese (Head of Division, ELS), and Glenda Quintini (Senior Economist, ELS). All remaining mistakes are solely the responsibility of the authors.

The authors are especially grateful to Glenda Quintini for her continuous guidance and feedback. The report further benefitted from comments and suggestions by Andrea Bassanini, Francesca Borgonovi, Stijn Broecke, Emanuele Ciani, Marie-Hélène Doumet, Alexandre Georgieff, Alexandre Lembcke, Fabio Manca, Carlo Menon, Lea Samek, Mariagrazia Squicciarini, Alexandra Tsvetkova, Annelore Verhagen, Wessel Vermeulen, Alison Weingarden, and participants to the OECD internal seminars and to the 2021 ESCOE Conference on Economic Measurement.

Abstract

This report presents a methodology to classify skill requirements in online job postings into a pre-existing expert-driven taxonomy of broader skill categories. The proposed approach uses a semi-supervised Machine Learning algorithm and relies on the actual meaning and definition of the skills. It allows for the classification of more than 17 000 unique skill keywords contained in the Burning Glass dataset into 61 categories. The outcome of the classification exercise is validated using O*NET information on skills by occupations, and by benchmarking the results of some empirical descriptive exercises against the existing literature. Compared to a manual classification, the proposed approach organises large amounts of skills information in an analytically tractable form, and with considerable savings in time and human resources.

Résumé

Ce rapport présente une méthodologie permettant de classifier les mots-clés se rapportant aux compétences mentionés dans les offres d'emploi en ligne dans une taxonomie préexistante conçue par des experts. L'approche proposée s'appuie sur un algorithme d'apprentissage automatique semi-supervisé qui utilise la définition et donc la signification des compétences. Elle permet de classer les 17 000 mots-clés de compétences qui apparaissent dans l'ensemble de la base de données Burning Glass Technologies en 61 catégories. Le résultat de l'exercice de classification est validé en comparant les compétences par occupation calculées grâce à O*NET et à l'information contenue dans la base de données Burning Glass Technologies et classifiée grâce à la méthode présentée. La reproduction de certains principaux résultats de la litérature permet de confimer la pertinence de l'exercice de classification. L'approche proposée permet donc de traiter une grande quantité de données sur les compétences de manière plus efficace en terme de temps et ressources humaines investis qu'une classification manuelle.

Table of contents

Tables

Figures

1 Introduction

1. Recent large databases of information derived from online job postings allow for rich analyses of labour market dynamics. The Burning Glass Technologies dataset (hereafter: Burning Glass data) contains a wealth of information on millions of job postings in several countries, including job and employer characteristics, and requirements in terms of education, professional experience, and skills. Using the list of skills demanded by employers, for instance, researchers have been able to analyse how skill requirements have evolved over time, and how they correlate with several measures of pay and firm performance (Deming and Kahn, 2018[1]; Hershbein and Kahn, 2018[2]).

2. Yet the number of distinct skills listed in the same dataset can be very large. While the rich information represents an important advantage over more traditional data sources, which do not usually contain such information or with high granularity and timeliness, the large number of listed skills also poses some challenges. First, the list contains several synonyms or closely related concepts that should be considered as such and analysed together, to avoid interpreting differences in terminology across sectors, places, or over time as true variation in skill requirements. Second, in many instances, such large number of skills cannot be easily or meaningfully described if not grouped in an appropriate way. These aspects stress the necessity to reduce the dimensionality of the skill information in the Burning Glass dataset to facilitate analysis. Because of the sheer number of unique skill keywords in Burning Glass, and the continuously updated dataset, a manual classification was ruled out as a viable solution.

3. The present study proposes an original approach to reduce the dimensionality of the skill information contained in the Burning Glass dataset. It does so by classifying the approximately 17 000 different skills appearing in Burning Glass data for Australia, Canada, New Zealand, Singapore, the United Kingdom and the United States into a pre-existing skill taxonomy based on the skill's meaning or definition. Instead of a manual classification, this study proposes a semi-supervised machine learning approach that produces an automatic classification of skills into the taxonomy's broader categories. The approach builds on BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art algorithm recently published by researchers at Google AI Language, which is trained on a large corpus of text to "understand" the English language. In the present context, the model is further trained (fine-tuned) for a specific task, i.e. the classification of skills keywords. The proposed approach addresses both issues discussed above: it accounts for the existence of synonyms, and classifies the long and time-changing list of Burning Glass skills into a final taxonomy that is stable over time. This stability in structure and the mutually exclusive nature of the taxonomy's categories are especially important to simplify empirical analysis. Furthermore, the algorithm can be launched every time new data are made available, and the reported new skills need to be classified – an exercise that would be potentially very time consuming if performed manually.

4. The proposed list of broader categories in which the skills will be classified (the *taxonomy*) largely builds on existing taxonomies, which were developed and validated by labour market and education experts and have been widely used by policy-makers and statistical agencies. A natural candidate to design such final taxonomy is the skill categorisation of the O*NET database. O*NET is a publicly available online database that, for each occupation, provides definition and main characteristics, including the skills, knowledge and abilities required to perform the job. This detailed information on skills use in occupations is organised under a clear hierarchical structure, which is frequently used for policy-making. The present study therefore makes extensive use of O*NET's existing information on skills, and complements it with

other information where O*NET is not sufficiently detailed to classify Burning Glass requirements (e.g. digital skills).

5. The outcome of the procedure is the *classification* of the 17 331 skills keywords contained in the original Burning Glass data onto a taxonomy of 61 skill categories. Because of the supervised nature of the exercise, it is possible to compute the model's accuracy, i.e., the percentage of correct predictions of the model evaluated against a test set, or a subset of the skills that were manually classified by the researchers. In the present exercise, accuracy ranges between 74% and 85%. These figures are much higher than accuracies obtained by randomly classifying the skills (2%) and mirror the degree of agreement reached between humans during the manual allocation of a sub-sample of skills. Indeed disagreements among researchers can happen in light of the intrinsic difficulty of classifying vague, incongruous or ill-defined skill keywords, hence lower accuracy than 100% should be expected.

6. As a result of the classification, the most populated categories are "Medicine and Dentistry", "Management of Financial Resources", and "Biology**",** which partially reflects the high frequency of health-related job postings in the Burning Glass database, and their propensity to include very specific skill requirements. Conversely, skills relating to abilities (as per O*NET definition) are very rare, possibly because they are implicit and are not clearly stated in the job advertisement in the first place.

7. The last section of this study displays some applications of the classification to the main Burning Glass datasets, with the aim of further validating the classification. The underlying assumption is that sizeable errors in the classification would reflect in empirical results that are inconsistent with the literature. A first validation exercise compares, for each occupation, the frequency of the classified Burning Glass skills with data on the importance of skill requirements as reported in O*NET. While the association is not expected to be perfect (e.g. due to difference between online and overall labour demand), the correlation is 52%, positive, highly significant, and robust to controlling for different levels of occupational aggregation.

8. A second set of validation exercises, instead, describes patterns of skill demand from online job postings, and attempts a comparison with results in the relevant literature. A first exercise looks at the variance in skill requirements, and finds that employer specificities and other unobservable factors explain a much larger share of the variance than occupation, education or experience attached to the job postings, which is consistent with findings in Deming and Kahn (2018[1]). A second exercise estimates the difference in the probability of requiring a certain skill if the job posting is opened in a digital intensive *vs* less digital intensive sector. For communication, digital, cognitive and managerial skills (such as resource management skills), online demand is higher in digital intensive than less digital intensive sectors, while the opposite is true for production or physical skills. These results are also in line with findings by Grundke et al. (2018[3]), computed using data on skill use on the job from the OECD Survey of Adult Skills. Lastly, industry-specific skills, but also cognitive, science, social and management skills are associated with relatively high posted wages, irrespective of the advertised sector, occupation or geographical location. Conversely, postings requiring mainly physical skills, skills most related to production or business processes, language skills and attitudes offer on average lower starting wages. These estimated wage returns are consistent with results from the existing literature (Deming and Kahn, 2018[1]; Hanushek et al., 2015[4]; Grinis, 2019[5]). From these validation exercises, we conclude that the methodology is suitable to classify a large amount of data while avoiding an extremely time-consuming and onerous manual task.

9. In the remainder of this study, Section 2 discusses the related literature, both the economic one using Burning Glass skill data and the data science one presenting Natural Language Processing models for text classification. Section 3 describes the dataset used and the final taxonomy based on O*NET. Section 4 introduces the methodology developed to map Burning Glass skills onto the final taxonomy, and Section 5 discusses the results of the classification exercise, including model's accuracy. Section 6 shows the results of several validation exercises comparing the results of the classification exercise with other expert-driven and data-driven classifications, as well as some descriptive statistics using the classification. Section 7 concludes.

2 Related studies

Leveraging on the skill information contained in Burning Glass

10. Several studies have used Burning Glass data to analyse skill dynamics in job postings. Burning Glass itself exploits this information in a number of policy reports, for instance to identify the fastest-growing occupations and skills in the U.S. job market, to analyse demand for IT skills in non-tech industries, or to study the demand for digital skills in the UK (Burning Glass Technologies, 2019[6]; Burning Glass Technologies, 2019[7]; Burning Glass Technologies, 2019[8]). The same data are also used in academic research about employers' skills demand: see for instance Börner et al. (2018[9]), Dorrer (2014[10]) and (Beblavý, Fabo and Lenearts, 2016[11]) for a focus on IT skills. Other researchers study the impact of the Great Recession, and highlight that skills requirements in job vacancy ads increased in locations particularly affected by the crisis (Hershbein and Kahn, 2018[2]; Modestino, Shoag and Ballance, 2019[12]). The impact was more pronounced for routine-cognitive occupations such as clerical, administrative, and sales occupations (Hershbein and Kahn, 2018[2]), and in high-wage cities (Blair and Deming, 2020[13]). Furthermore, the effects were persistent among high-skill jobs while increases in skill requirements in job vacancy ads for middle-skill and low-skill occupations were either temporary or non-existent (Burke et al., 2019[14]), or tended to reverse as the labour market recovered (Modestino, Shoag and Ballance, 2016[15]).

11. In exploring the causes of these changes in skill requirements, Kuhn, Luck and Mansour (2018[16]) show that firms affected by offshoring-related layoffs increase their demand for soft skills, such as communication and teamwork. Dillender and Forsythe (2019[17]) argue that an increase in skill requirements can be linked to the adoption of new software technology at the job-title level. Campello, Gao and Xu (2019[18]) demonstrate negative effects of local personal income taxes on skill requirements, and interpret them as a tax-induced "brain-drain" having a detrimental effect on the skill composition of local labour markets.

12. In turn, these changing job skill requirements have a profound impact on the dynamics of the labour market and on the wider economy. For instance, job openings including STEM requirements take longer to fill (Rothwell, 2014[19]). Earnings dynamics during the career are also affected: the earnings premium for college graduates majoring in technology-intensive subjects declines rapidly (Deming and Noray, 2020[20]). Consequently, these individuals move out of faster-changing occupations as they gain experience. Deming and Kahn (2018[1]) also present evidence of a positive correlation between employer demand for cognitive and social skills, and measures of pay and firm performance.

13. To conduct these empirical investigations, researchers do not use the skill information as contained in the Burning Glass dataset, but rather rely on various methods to reduce the dimensionality of the information. First, some studies use requirements in terms of education and experience as indirect measures of skill requirements, therefore discarding a lot of potentially useful information (Blair and Deming, 2020[13]; Modestino, Shoag and Ballance, 2019[12]).

14. A number of studies exploit the skill taxonomy built by Burning Glass, mapping the approximately 17 000 distinct skill keywords onto 1 200 "clusters" and 28 "families". This is the case for instance of Burke et al. (2019[14]), Dorrer (2014[10]) or Modestino, Shoag and Ballance (2016[15]). However, the methodology used by Burning Glass to build their taxonomy is not explained clearly, and the taxonomy itself tends to

reflect industry categories rather than actual skills. Furthermore, in some instances, the same variable contains what should be considered different levels of a skill hierarchy; for instance, a Burning Glass skill cluster variable lists both "Secretarial" and "Record Keeping" skills, therefore implicitly suggesting that these two concepts are at the same hierarchical level of a skill taxonomy.

15. Alternative approaches to reduce the dimensionality of Burning Glass skills include Börner et al. (2018[9]), who rely on a manual tagging of hard *vs* soft skills for approximately 3 000 skills that appear in data science and data engineering jobs. A time-consuming endeavour, this requires further investments in extending the tagging every time the underlying Burning Glass data introduce a new skill. Recent OECD work by Brüning and Mangeol (2020[21]) manually classify the skills into four main categories: cognitive, socio-emotional, technical transferable, or technical job-specific, where the first three categories include skills potentially applicable to a large range of jobs and therefore defined as "transferable". The classification is applied to skills in job postings for the United States in 2018 that have non-missing occupation information and require a higher education degree.

16. Other works use a restricted number of categories to identify particular skills (Beblavý, Fabo and Lenearts, 2016[11]; Campello, Gao and Xu, 2019[18]; Deming and Kahn, 2018[1]; Deming and Noray, 2020[20]; Hershbein and Kahn, 2018[2]; Kuhn, Luck and Mansour, 2018[16]). In these cases, the authors specify a number of "final" categories (ranging from 2 to 14), as well as the different skill keywords which fall into each category ex-ante. For instance, (Deming and Kahn, 2018[1]) consider ten categories: cognitive, social, character, writing, customer service, project management, people management, financial, budgeting, computer, and software skills. Job vacancies that contain keywords and phrases such as "problem solving," "research," and "analytical" are considered as requiring cognitive skills, and similar definitions are developed for the other nine categories. The authors therefore must identify the categories that are important in the labour market and the list of accepted keywords ex-ante. This, however, is only feasible when the number of skills categories is small and therefore, under this approach, the proposed list of skill categories is not exhaustive and not all skills contained in Burning Glass data can be classified. Other studies link Burning Glass skill information with expert-driven skill classifications. For instance, to calculate the education, training, and skill requirements of job ads and classify them into STEM or non-STEM openings, Rothwell (2014[19]) matches the Burning Glass dataset with data from the U.S. Bureau of Labor Statistics and from O*NET, based on occupational codes.

17. Finally, a nascent strand of literature implements purely data-driven Machine Learning algorithms to classify Burning Glass skills into a smaller number of categories, as in e.g. Djumalieva and Sleeman (2018[22]). The authors model skills and their relationship using a network graph, where skills are represented as nodes and are connected with lines according to whether and how often they co-occur in job postings. The relationship between two skills (co-occurrence) is measured using both simple pairwise co-occurrences (number of times two skills appear in the same advert), as well as their shared context. After building the network graph, similar skills are grouped together using clustering techniques. This approach allows identifying groups of similar skills beyond pairwise connections, but also of transversal skills. However, the authors' approach also presents some important limitations. First, it is not clear how new data waves can be integrated, nor how this will modify results (in particular, the number and nature of final groupings). Second, the labelling of different clusters is challenging: according to the authors themselves, a first attempt to create data-driven labels failed, as it did not produce names that were representative of the whole content of each respective category, and manual labelling is very time consuming. Another attempt to group skills based on their co-occurrences, with the same caveats as above, is presented in Dawson et al. (2019[23]). They first compute the relative importance of a skill in a job ad, and then identify skills as similar when they are found to be relatively important in the same job ads.

Natural Language Processing techniques for classification

18. Text classification, i.e. the transformation of unstructured textual data (documents, books, reports, etc.) in a structured format, is a long-standing problem in the field of data science. Recent breakthroughs in Natural Language Processing (NLP) have allowed significant progress in this matter and have permitted a number of important real-life applications. Text classification usually involves the following steps: Feature Extraction, Dimensionality Reduction, and finally Classification (Kowsari et al., 2020[24]).

19. The first step, *feature extraction*, is the process of transforming unstructured textual data in a mathematical object (e.g. a vector) that can be used later by a classifier. A preliminary cleaning step is often needed, including stop words removal, noise removal, spelling correction, capitalisation, tokenisation, and/or lemmatisation. The data (depending on the application: characters, groups of characters also known as n-grams, words, sentences or paragraphs) are then transformed into vectors. Popular methods to achieve this transformation include Bag of Words (BoW), TF-IDF Vectorisation (Jones, 1972[25]) and Word embeddings (Mikolov et al., 2013[26]; Pennington, Socher and Manning, 2014[27]). Bag of Words models represent words as one-hot vectors (consisting of 0s in all cells with the exception of a single 1 in a cell used uniquely to identify the word). This is the simplest technique but it presents a major drawback: two related words may not be identified as such because they can be attributed radically different vectors. More recently, TF-IDF and, to a greater extent, word embedding techniques, have delivered unprecedented improvements in text classification by representing words as dense vectors (where most elements are non-zero), as they allow for the detection of finer relationships between words. Still, these methods fail to take into account the context in which a word can be used and thus showed poor performance for words that acquire different meanings depending on context. In the late 2010s, major advances were achieved thanks to the development of contextualised representations methods (Peters et al., 2018[28]). With these techniques, words are not attributed a single vector and their representation depends on the sentence in which they are embedded.

20. The second step, *dimensionality reduction*, is optional. Its aim is to map the word representations into a vector space with a lower dimensionality, to simplify the classification at a later stage. Traditional techniques include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), T-distributed Stochastic Neighbour Embedding (t-SNE) or other projection algorithms. More recently, Autoencoders (Goodfellow, Bengio and Courville, 2016[29]) have achieved great success in this task. Autoencoders are Artificial Neural Networks (ANNs) transforming a list of input vectors into a list of output vectors. Dimensionality reduction occurs when the size of the output list is smaller than the size of the input list.

21. The *classification* part consists in mapping the vector representation of a text (characters, n-grams, words, sentences or paragraphs) into one of several predetermined categories.[1] This can be achieved with classification algorithms that are not specific to NLP: logistic regression, k-nearest neighbours, support vector machines (SVM), naïve Bayes, decision trees, etc. The most recent state-of-the-art models rely on Deep Neural Networks (DNNs) with a final softmax classifier. In these models, the dimensionality reduction and the classification parts are often merged into a single step: the model takes a pre-processed sentence as an input (with a list of vectors or contextualised representations for the different parts of the sentence) and produces one predicted category. Successful implementations include Recurrent Neural Networks (RNNs), especially Long-Short Term Memory networks (Hochreiter and Schmidhuber, 1997[30]; Graves and Schmidhuber, 2005[31]) and Gated Recurrent Units (Chung et al., 2014[32]).

22. Since 2018, the performance of text classification models has dramatically improved with the introduction of language models. These models implement the different steps described above, from

[1] The process of assembling text into different groups and giving these groups a name or concept ex-post is called *clustering* and uses radically different algorithms. These techniques fall outside the scope of this study and are therefore not discussed here.

feature extraction to classification, within a single architecture relying on Artificial Neural Networks. The models' weights are trained on huge corpuses of textual data to acquire a general knowledge of language. ULMFiT (Howard and Ruder, 2018[33]) then BERT (Devlin et al., 2018[34]) are the two milestones models in this field. Other, more recent, models have improved their performance: RoBERTa (Liu et al., 2019[35]) proposed by Facebook AI in 2019 and building on BERT, XLNet (Yang et al., 2019[36]) and GPT2 (Radford et al., 2019[37]). However, for the purpose of this project, the improvements brought about by these latter models are marginal, and come at the cost of utter complexity.

3 An overview of the data

Skill information in Burning Glass data

23. The present work relies on data transmitted by Burning Glass in 2019. The dataset includes information on more than 200 million job ads[2] gathered online between 2012 and 2018 across six English-speaking countries: Australia (AUS), Canada (CAN), New Zealand (NZL), Singapore (SGP), United Kingdom (GBR) and United States (USA). Data for 2007, 2010 and 2011 are also available, but for the United States only. Data for Australia and New Zealand are provided in a single file, hence the statistics below are offered for both countries jointly.

24. The Burning Glass dataset contains standardised information retrieved from job postings using more than 45 000 online sources. It includes job characteristics such as detailed industry and occupation codes, location, posting date, name of the employer, and requirements in terms of education, professional experience, and skills. This great level of detail allows analysing job requirements within - rather than only across - occupations, sectors, and locations. Such granular analyses usually cannot be performed with traditional data sources. Furthermore, since Burning Glass data are updated more frequently than administrative datasets and surveys and often in real or quasi-real time, they also allow for an earlier detection of emerging trends than previously possible.

25. A number of existing studies describe the representativeness of Burning Glass data. Carnevale, Jayasundera and Repnikov (2014[38]) show that, for the United States in 2006-2013, the aggregate number of job postings in Burning Glass strongly correlates with the number of job openings reported in the Bureau of Labor Statistics Job Openings and Labor Turnover Survey (JOLTS). If Burning Glass data overrepresent openings for high-skilled jobs, this feature is constant over time (Deming and Kahn, 2018[1]). Carnevale, Jayasundera and Repnikov (2014[38]) estimate that, in the U.S. files, state, city, occupation title, major occupation group, skills, and education are correctly reported for at least 80% of observations with non-missing information, while accuracy is lower for minor occupation groups, and industry codes. Recent OECD work (Cammeraat and Squicciarini, 2021[39]) has also analysed the representativeness of Burning Glass data against official employment data at the occupational level, showing that for the period 2010-2018, for the majority of countries, Burning Glass data is of sufficiently good quality to conduct policy analyses. Representativeness concerns exist for Canada and New Zealand, but issues emerge mostly for the years prior to 2015 and representativeness has improved since.

26. To identify skills required to perform the job, Burning Glass analyses the text of each job vacancy. This information is processed and standardised, e.g. by removing duplicates, or by treating differences in spelling for the same skill, with the notable exception of British versus American English (e.g. “Organizational skills” in CAN and USA but “Organisational skills” in AUS, NZL, SGP and GBR). Skill keywords, however, may still include acronyms (e.g. “ADHD Tutoring”).

27. These keywords include skills in the sense which is commonly understood (e.g. “Analytical Skills”), but also knowledge (e.g. “Food Safety” or “Environmental Policy”) and abilities (e.g. “Detail-Oriented”).

[2] 217 461 987 more precisely.

While there is no easy way to sort Burning Glass skills into these three conceptual categories in an automated way, a first assessment suggests that skill keywords mostly capture knowledge areas.

28. In total, there are 17 511 different unique skill keywords (henceforth "skills") across all years and countries:

- 8 595 (49%) are common to all five countries or areas (ANZ, CAN, SGP, GBR, USA);
- 10 476 (60%) are common to at least four countries;
- 12 913 (74%) are common to at least three countries;
- 15 259 (87%) are common to at least two countries;
- 2 115 (12%) are unique to one country.

29. The fact that some skills are unique to one or few countries stems from two major factors. First, some skills are so rare that they appear only once or twice in the whole dataset, and consequently, in a single country and year. Second, several skills are country-specific for geographical or historical reasons (e.g. "Knowledge of Aboriginal Heritage Law" in ANZ or "Inuit Health" in CAN). As explained in the section below, the supervised methodology is robust to the rarity of skills: as long as rare skills are well defined, they can be well classified by the algorithm.

30. Likewise, cultural norms, features of the posting platform, or simply employers' choices and habits in writing may introduce differences across skills even when some keywords are actual synonyms, which may inflate the overall number of skills available in the dataset. While these differences cannot be identified ahead of processing the data, they will be factored in henceforth, when constructing the classification. In fact, the methodology proposed below aims precisely to classify synonyms under the same broader category.

31. Nonetheless, some job postings that Burning Glass retrieved online have no skill requirements information. Table 1. shows the proportion of all job ads for which skill requirement information is available, per country and per year. This proportion is particularly high in the US, where more than 98% of job ads contain skill requirements (except for the year 2018 where this proportion is slightly lower). The availability of skill information is the lowest and most heterogeneous in Canada: depending on the year, the percentage of observations with non-missing skill requirements ranges between 83% and 98%. In other countries, the share of observations with missing skill information is stable at around 10% across the years.

Table 1. Percentage of observations with non-missing skill requirements

	AUS-NZL	CAN	SGP	GBR	USA
2007					98.5
2010					98.3
2011					98.6
2012	93.0	97.7	88.9	91.6	98.6
2013	88.6	96.9	90.9	89.9	98.9
2014	90.0	87.3	90.2	90.5	98.9
2015	91.2	84.5	92.2	89.9	99.0
2016	90.3	83.4	93.8	89.2	99.1
2017	90.8	92.1	93.6	89.9	99.1
2018	90.8	87.4	93.6	89.7	95.0

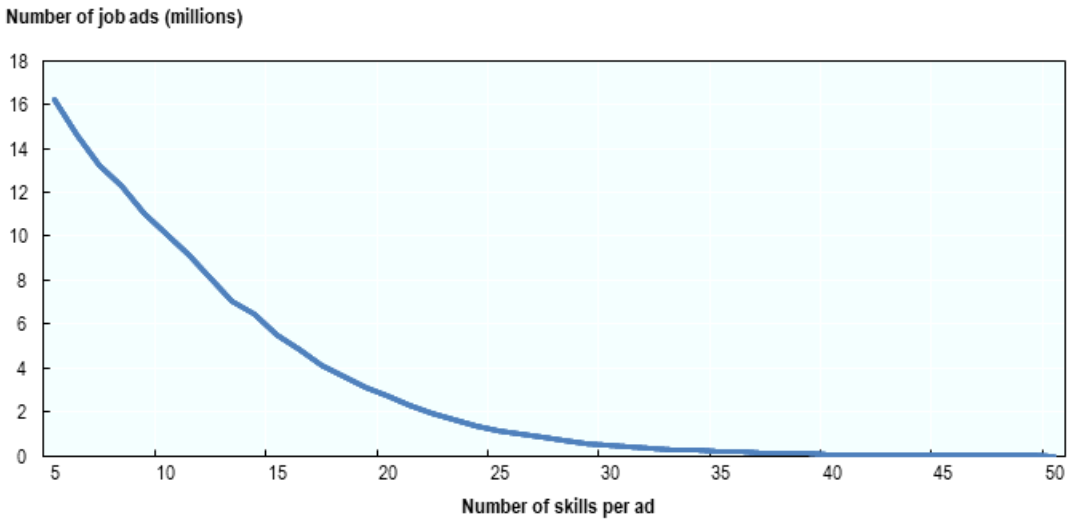
Source: OECD calculations based on Burning Glass data.

32. While it cannot be excluded that a lack of skill requirements in a posting is caused by the failure of Burning Glass's algorithm to retrieve such information from the text of the job advertisement, many job postings simply do not report any precise skill requirement. This is the case, for instance, of jobs for which requirements are implicitly conveyed by the job title or by the education requirements mentioned in the job ad. While the existence of implicit skills has important consequences for analysis, it does not affect the outcome of the taxonomy construction as approached in this study, insofar as only the skills contained in the Burning Glass dataset need to be classified.

33. Annex A.1 shows similar figures on job postings with non-missing skill information, broken down by occupation (SOC codes, truncated to the first 3 digits). For the vast majority of occupations, the share of posting without skill information is lower than 10%. Notable exceptions include occupations corresponding to codes 39-4 (funeral workers), 45-1 (supervisors of farming, fishing, and forestry workers), 45-2 (agricultural workers), and to a lesser extent codes 17-1 (architects) and 53-5 (water transportation workers).

34. Job ads that do list skills requirements contain on average 7.8 skills keywords (median equal to 5). Most job ads (>98%) require less than 20 skills (Figure 1), and a very small number of ads require more than 300 skills. This likely stems from errors in Burning Glass parsing technique, or from the fact that these postings are probably not unique job openings but rather collections of offers. As the classification approach in this study does not exploit additional information contained in job postings other than skill keywords and their meaning, these incongruences do not affect the outcome of the classification exercise. In the present work, therefore, all skills are included and treated equally.

Figure 1. Distribution of the number of job ads per number of skills required



Note: This graph shows data for jobs ads with 5 to 50 skill requirements, for the sake of readability. These represent 94% of job ads in the dataset.
Source: OECD calculations based on Burning Glass data.

35. Table 2 presents summary statistics of the dataset by country, focusing on skill requirements. The United States, and to a lesser extent the United Kingdom, account for the vast majority of job postings with at least one skill requirement. US-based offers also gather significantly more skills (16 048) and country-specific skills (1 497) than the others. The average number of skills per ad varies from 5.4 for Australia-New Zealand to 9.4 for Canada.

36.

Table 2. Summary statistics of the skills files in Burning Glass data

	AUS-NZL	CAN	SGP	GBR	USA
Number of job ads with at least one skill requirement	6 974 051	6 487 666	3 489 531	52 393 082	148 117 657
Number of distinct skills	11 608	12 979	10 881	13 238	16 048
Number of country-specific skills	75	20	283	240	1 497
Average number of skills per ad	5.4	9.4	7.4	5.7	8.6

Source: OECD calculations based on Burning Glass data.

The O*NET+ taxonomy

37. The starting point for the construction of a pre-set skills taxonomy was the O*NET database from the U.S. Department of Labor. The database was created in 1998, building on its predecessor the Dictionary of Occupational Titles (DOT), and is updated on a regular basis. O*NET contains a wealth of information on occupations, including the skills, knowledge and abilities needed to work in each of the almost 1 000 occupations. Most of the occupational information, including knowledge requirements, is collected from job incumbents through surveys. Skill and ability requirements are determined by occupational analysts, who take information on tasks, knowledge, vocational preparation, work activities and work styles from the incumbents' surveys as a starting point (Tsacoumis and Willison, 2010[40]). In the O*NET database, the skill, ability and knowledge requirements of occupations are measured both in terms of importance and level. The former indicates whether the particular skill, knowledge or ability is important to perform the job. The latter indicates the level of mastery or proficiency in that skill, knowledge, or ability needed for the job.[3] The latest version of the database contains 33 knowledge types, 35 skills and 52 abilities. O*NET defines skills as "developed capacities that facilitate learning or the more rapid acquisition of knowledge", abilities as "enduring attributes of the individual that influence performance", and knowledge as "organized sets of principles and facts applying in general domains". While the distinction between skills and abilities is rather subtle, the difference compared to knowledge is clearer. For example, O*NET contains both Mathematics Knowledge and Mathematics Skills, the former being defined as "knowledge of arithmetic, algebra, geometry, calculus, statistics, and their applications" and the latter as the capability to use mathematics knowledge to solve problem. All of the O*NET skills, abilities and knowledge types have been grouped into more aggregate categories by the authors of this study, as reported in Table B.1.

38. As O*NET is developed by the Bureau of Labor Statistics in the United States, it is geared towards the occupational content of jobs in the labour market in the United States. Despite this, O*NET has been regularly used for the analysis of countries other than the United States. The assumption that skill measures from one country can be generalised to other countries has been tested and largely holds (Cedefop, 2013[41]; Koucký, Kovařovic and Lepič, 2012[42]; Handel, 2012[43]). Handel (2012[43]), for example, finds that occupational titles refer to very similar activities and skill demands across different countries. Specifically, high correlations between O*NET scores and parallel measures from the European Social Survey, EU Labour Force Survey, Canadian skill scores, the International Social Survey Program, and the UK Skill Survey are found, with average correlations of 0.80. Other than a handful of occupational skill requirements that exhibited significant cross-national variation (i.e. prior experience required, training required, and job learning times), Handel (2012[43]) found that most skill scores can be generalised to other countries with a reasonable degree of confidence. A caveat should, however, be raised about the use of

[3] Ratings on Level are collected on a 0-7 scale, and ratings on Importance are collected on a 1 ("Not Important") - 5 ("Extremely Important") scale.

O*NET to describe skills and tasks of occupations in low-income countries, as these could differ significantly in terms of technology and regulatory context compared to the United States.

39. The O*NET information on skills, knowledge and ability requirements has been used extensively in labour market research. For example, Deming (2017[44]) uses the database to measure the extent to which occupations use non-routine analytical tasks, service tasks, and social skills. Consoli et al. (2019[45]) use O*NET skills, abilities and knowledge requirements to determine the routine intensity of occupations, building on earlier work carried out with O*NET's predecessor DOT (e.g. Goos, Manning and Salomons (2014[46])). The OECD Skills for Jobs Database uses O*NET to map occupational imbalances into shortages and surpluses of skills, abilities and knowledge types (OECD, 2017[47]).

40. Given the wide use of the O*NET database, its long history, and the fact that it is built on expert views, the skills, knowledge types and abilities, and the hierarchical structure of those, serve as an natural starting point for the creation of a taxonomy to use for classifying the Burning Glass skill keywords.[4] However, this database also has some shortcomings.

41. One of the key gaps in the O*NET skills, knowledge and abilities classification is its limited inclusion of digital skills. In light of the growing importance of these skills, and the frequency with which they appear in online vacancies, including more detailed digital skills would increase the usefulness of the taxonomy for the purpose of the classification exercise. O*NET's digital skills are limited to: i) knowledge of computers and electronics, and ii) programming skills. In order to bring in a more detailed digital skills component to the classification, these two components were replaced by a set of six broad digital skills: Digital content creation; Digital data processing; ICT safety, networks and servers; Office tools and collaboration; Software; and Computer programming. These groups were created based on the ESCO (European Skills/Competences, Qualifications and Occupations) classification that includes similar categories.[5]

42. In addition to these digital skills, the ESCO classification is also used to add a set of attitudes and values to the taxonomy. These are not part of skills, knowledge types and abilities in O*NET[6], but feature frequently among skill keywords in job vacancies as reported by Burning Glass. While the ESCO classification includes 3 "values" and 16 "attitudes", for the purpose of the present study these categories are grouped into one "values" group and three "attitudes" groups.

43. This revised O*NET classification, with expanded digital skill categories and the additional "values" and "attitudes" categories, serves as the baseline taxonomy for the mapping of Burning Glass skill keywords in this study. When using this revised O*NET classification for the training set (see section 4), a few practical challenges emerged. Some of the knowledge, skills and abilities groups actually refer to very similar concepts. For example, O*NET contains information on the requirements of mathematics

[4] The lowest level of disaggregation of O*NET skills and knowledge types and the intermediate level for abilities is used as the baseline for the creating of the O*NET+ taxonomy. Given that abilities have three levels of disaggregation, while skills and knowledge only have two, the same level of disaggregation is used for the three types of requirements. Some ability categories are dropped, as they are sufficiently captured by other skills categories or have limited relevance in the classification exercise (i.e. memory, perceptual abilities, spatial abilities, attentiveness, and verbal abilities).

[5] ESCO contains the following subgroups within its digital skills category: programming computer systems; setting up and protecting computer systems; accessing and analysing digital data; using digital tools for collaboration, content creation and problem solving; using digital tools to control machinery. The latter category has not been kept in the taxonomy exercise, as it fits better in some machine operation-related skills, knowledge and abilities groups. By contrast, a "software" group was added in light of the high frequency of specific software keywords in the Burning Glass data.

[6] Nonetheless, the O*NET database contains information on work styles, which are similar to the values and attitudes from the ESCO classification.

knowledge, mathematics skills, and mathematical reasoning. While conceptually these are all different from one another, this distinction is hard to capture in real-life vacancy data. This points towards an important conceptual difference between the O*NET database and the taxonomy intended for classifying the Burning Glass skills: the skills, knowledge types and abilities in O*NET can be partially overlapping, whereas the groups in the desired taxonomy to use for the Burning Glass data need to be mutually exclusive. Put differently, a Burning Glass skill keyword could be classified in multiple O*NET categories as the keyword encompasses skills, abilities and knowledge in similar domains but, for the purpose of the machine learning classification exercise in this study, each keyword should only be allocated to one specific category in the taxonomy. This issue was particularly evident for skills keywords related to production and technology, for example, which could both refer to mechanics knowledge and installation or repairing skills, but also to certain skills related to business and management, which could relate to either economics and accounting knowledge or financial resource management skills. To facilitate the classification exercise, these O*NET categories were merged. The methodological approach and machine-learning algorithm proposed in the next sections are a consequence of the choice of classifying each skill in one category, and one category only.

44. The Burning Glass data also contain various skill keywords that refer to specific industry knowledge, such as automotive industry knowledge. Since such skill requirements in job ads generally refer to a broad set of underlying skills and knowledge requirements, these keywords are hard to classify into one single group in the taxonomy and are therefore all put under a separate “industry knowledge” group.

45. The final taxonomy, referred to as the O*NET+ taxonomy, contains 61 categories (see Table 3 for the full list). Three of those groups are coded separately, instead of relying on machine learning. Two of these categories refer to languages, one for “English” and the other for all non-English “foreign” languages. The latter is filled by calling an exhaustive list of all spoken languages (and broken down by whether it is the local language in the region/country or not) which currently contains 32 languages. A third category is filled automatically with Burning Glass skill keywords containing the “industry knowledge” phrase (297 keywords in the current version of the dataset). Finally, all skill keywords containing the phrase “Working with Patient and/or Condition: XXX” (416 skill keywords) are coded to be classified into the “Medicine and Dentistry” category. As a result, the Machine Learning model has to classify a skill into one of 58 categories.

46. While 61 categories are easier to handle in analysis than 17 000 skills, some research may need even more aggregate categories. A hierarchy is therefore proposed, where the 61 categories are grouped into 16 broader categories. Some of these broader categories are composed exclusively of skills, knowledge types or abilities, whereas others mix these different concepts when skills, abilities and knowledge pertain to the same or very similar domains. This is the case, for example, of the “production and technology” higher-level category, which contains both “service orientation” (a skill) and “customer and personal service” (knowledge).

Table 3. The O*NET+ taxonomy

Broad category	Category label	Broad category	Category label
Attitudes	Adaptability/resilience	Cognitive Skills	Originality
	Motivation/commitment		Quantitative Abilities
	Self-management/rigour		Reasoning and Problem-solving
	Work Ethics		Learning
Arts and Humanities	Fine Arts	Communication	Active Listening
	History and Archaeology		Reading Comprehension
	Philosophy and Theology		Speaking
Business Processes	Clerical		Writing
	Sales and Marketing		Communications and Media
	Customer and Personal Service	Digital	Office Tools and Collaboration Software
Production and Technology	Telecommunications		Digital Content Creation
	Building and Construction		Digital Data Processing
	Engineering, Mechanics and Technology		ICT Safety, Networks and Servers
	Design		Computer Programming
	Food Production		Web Development and Cloud Technologies
	Production and Processing	Resource Management	Time Management
	Transportation		Management of Material Resources
	Equipment Selection		Management of Financial Resources
	Quality Control Analysis		Management of Personnel Resources
	Installation and Maintenance		Administration and Management
Medicine	Medicine and Dentistry	Social Skills	Coordination
	Psychology, Therapy, Counselling		Persuasion and Negotiation
Law and Public Safety	Law and Government		Social Perceptiveness
	Public Safety and Security		Judgment and Decision Making
Science	Biology	Training and Education	Training and Education
	Chemistry	Physical Skills	Psychomotor Abilities
	Geography		Auditory and Speech Abilities
	Physics		Visual Abilities
	Sociology and Anthropology		Physical Abilities
Industry Specific Knowledge	Industry Specific Knowledge	Languages	Local language
			Foreign language

Source: OECD elaboration on O*NET and ESCO.

4 Mapping skills onto the O*NET+ taxonomy: the methodology

47. A supervised learning approach is used to map Burning Glass skills onto the predetermined O*NET+ categories, based on the skills' definition. The classification is done using BERT, a Natural Language Processing model that can perform sentence classification. BERT is a deep learning algorithm that contains mathematical representations of words (vectors), learned from external sets of texts. It is able to summarise the vectors corresponding to different words in a sentence (in the present case, in the skill's definitions) into a single vector, to perform the classification of this sentence. The algorithm relies on the definitions of skills and on the manual classification of 500 examples from which it learns how to associate categories to skills (training set). The skills' definitions are extracted from ESCO and Wikipedia, as specified further here below.

The BERT algorithm

Bidirectional Encoder Representations from Transformers

48. BERT (Bidirectional Encoder Representations from Transformers) is a language model developed by Google AI Language and published in 2018 (Devlin et al., 2018[34]). In Natural Language Processing (NLP), language models are neural network architectures that are able to process textual information, especially sentences. To that end, most recent language models build a mathematical representation for each word: a vector called an **embedding**. The rules to construct and process embeddings through the network (sometimes called the encoding process) are what define the language model.

49. A single word, however, can have different meanings depending on the context in which it is used (e.g. the prison cell, the biological cell, the cell phone, etc.), and the vector representation of a word has to evolve according to the context in which it is read. Therefore, for a given sentence, BERT produces different successive iterations of embeddings (twelve iterations in total[7]) where each new iteration takes into account all the embeddings from the previous iteration. When BERT computes a new iteration, i.e. a more specific and relevant representation for a word in the sentence, the choice of other words to consider is given by the so-called **Transformer** (Vaswani et al., 2017[48])**.** The Transformer is able to look at different parts of a sentence with different focus, a process called **attention**.[8]

[7] The number 12 is a trade-off between model size and efficiency. Empirically, more layers tend to offer a better understanding.

[8] The adjective "bidirectional" in the BERT acronym refers to the fact that the attention is drawn both to the left and right of the sentence, contrarily to other models that only look at the preceding words (such as Recurrent Neural Networks or GPT).

50. Transformers gather knowledge about the words that are most appropriate through a process of trial and error. For each iteration, the Transformer takes an existing well-constructed sentence, masks randomly some words, and asks BERT to recover the masked words (Liu et al., 2019[35]).[9] BERT predicts a probability for each word in its vocabulary to be the masked one. Depending on the precision of the guess, an error is computed and used to improve the model's ability to recover a masked word in the next iteration. Since each improvement is marginal, a tremendous amount of data is needed to train the whole model. Indeed, BERT has been trained on millions of sentences extracted from BookCorpus and Wikipedia (respectively 800 and 2 500 million words). BERT thus acquired a general knowledge of language that is used and summarised in embeddings. This process is called a **semi-supervised learning step** and BERT is a pre-trained model.

51. Several versions of BERT are available. This study uses BERT (Base), which is the language model that provides the best compromise between accessibility, performance, and widespread use in the literature. Annex C discusses other versions of BERT and other recent language models, and justifies the choice for BERT (Base) in detail.

BERT for sentence classification: a synthesis of the approach

52. The present study is aimed at one precise downstream application, i.e. the classification of Burning Glass skills into a small number of encompassing categories (the 58 O*NET+ categories presented above). The version of BERT used in this study does not work on the skill itself as a word, but on its definition, which is then transformed into an embedding to be classified. This adapted version of BERT is often referred to as "BERT for sentence classification". Starting from the assumption that the definition of the skill contains all the sufficient information needed for the skill's classification, the classification exercise consists in mapping each skill's "embedded" definition into one of the 58 O*NET+ categories.[10]

53. To achieve this goal, the definition of each Burning Glass skill is first processed through BERT to obtain the list of corresponding embeddings. Calling N the number of words in the definition of each Burning Glass skill, and K the dimension of one embedding vector[11], applying the BERT algorithm yields N vectors[12] with K components for each skill keyword, where N can vary between skills (different skills have definitions of different lengths). BERT then summarises the information contained in the sentence, hence in the N embeddings into a single vector. To do so, an obvious solution would be to take the average of all N vectors. However, the developers of BERT proposed a different approach that substantially improves the algorithm's performance over using a mean vector: adding an additional fictitious word, called **CLS token**, which at the end of the encoding procedure represents the entirety of the information contained in the sentence.[13]

[9] BERT was also trained with a second objective of predicting if two sentences follow each other or not. Later work showed that this objective can be dropped without affecting the results ("RoBERTa: A Robustly Optimized BERT Pretraining Approach" - Liu et al., 2019).

[10] The approach chosen to build the present classification asks each skill to be classified in one category only; assigning multiple categories to each skill based on the 58 values of the vector is therefore inconsistent with the methodology proposed here. Researchers interested in assigning multiple plausible categories to a skill should investigate the use of different methodologies (multiclass labelling).

[11] In BERT, $K = 768$.

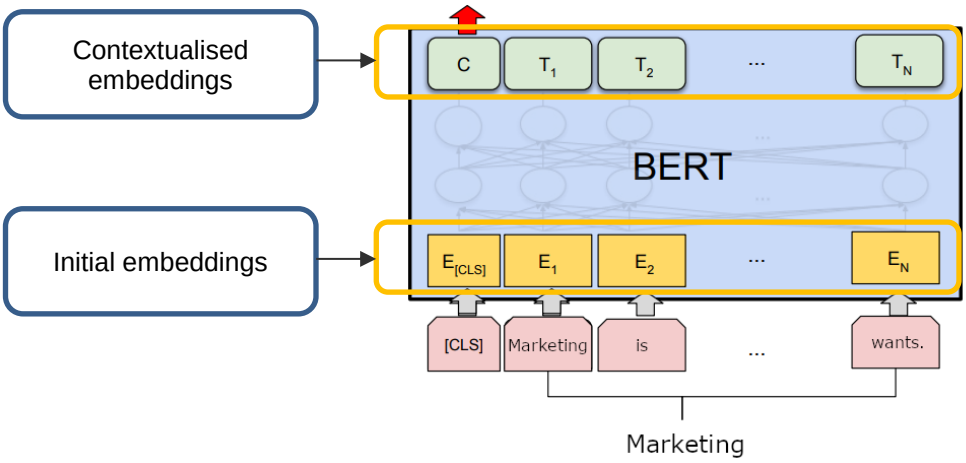
[12] In fact, more than *N*, because of tokenisation - see Annex C.

[13] Using an average vector of the different embeddings has the limitation that less relevant words in the definition (e.g. linking words and connectors such as "a", "the", or "of") obtain the same weight as more relevant words. This dilutes the informative power of relevant words when embedded.

54. Figure 2 proposes a graphical, simplified representation of the process as described so far. In this example, the algorithm must classify the skill "Marketing" into one of the 58 outstanding categories of the ONET+ taxonomy. The definition of "Marketing" is composed by various words, in pink ("marketing", "is", "wants", etc., including the fictitious CLS token). The initial embeddings of these words (in yellow) are elaborated by BERT through 12 layers of processing, to add contextual information to each word embedding. This process yields the contextualised embeddings (in green, at the top of the box), including the embedding of the CLS token, identified as "C" in the figure.

$$C = \begin{pmatrix} c_1 \\ \vdots \\ c_K \end{pmatrix}$$

Figure 2. The keyword "Marketing" and its definition processed through BERT



Source: OECD elaboration on Devlin et al. (2018[34]).

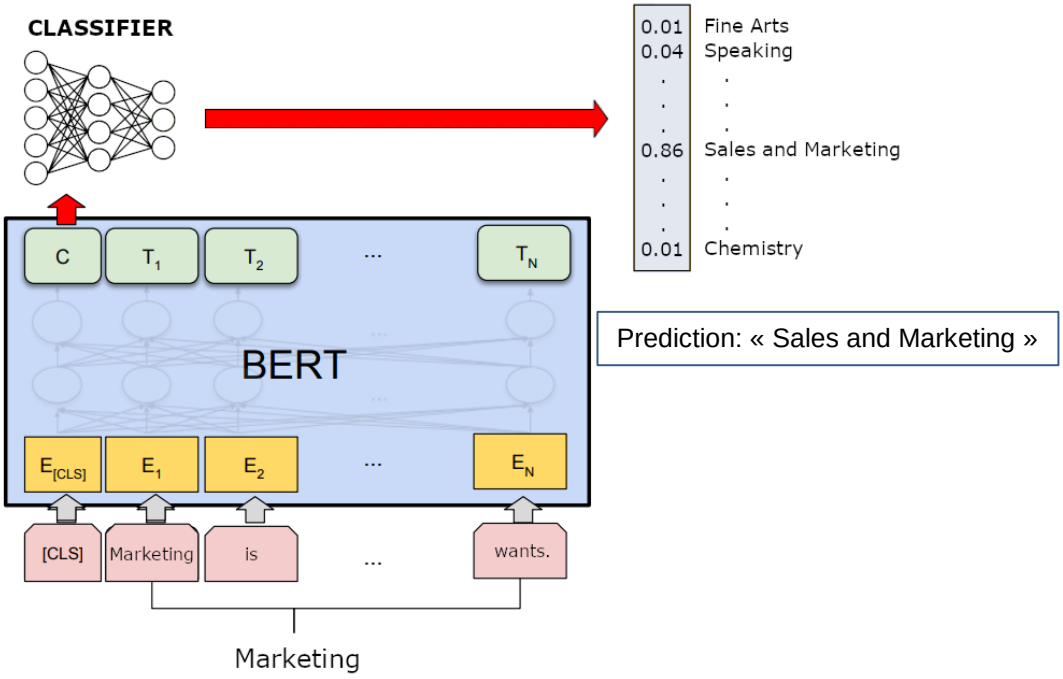
55. Having obtained a single vector for the definition of every skill, this must be associated to one of the 58 O*NET+ classes. The standard method proposed by BERT for sentence classification is pictured in Figure 3 in a simplified form.[14] It relies on a function mapping a real vector into a positive vector whose 58 components sum to 1 (one component for each O*NET+ class). It is traditionally called "softmax function" in the literature and is defined by:

$$\sigma(c)_j = \frac{e^{c_j}}{\sum_{k=1}^{K} e^{c_k}}$$

56. for all *j=1,…, 58.* The output of the softmax function is thus a vector of 58 values ranging from 0 to 1, a transformation of the original vector corresponding to the skill's definition. After the optimisation step (see next section), the highest of the 58 values within this vector corresponds to the category where the model predicts the skill should be classified. For example, in Figure 3 the highest value is 0.86 and the corresponding category is "Sales and Marketing".

[14] More precisely, the methodology relies on a one-layer feed-forward neural network with a softmax output, where the term "one-layer" highlights that the approach uses only one matrix operation that maps the 768-component vector into a 58-component vector, and "feed-forward" distinguishes this approach to others previously used in the literature, such as recurrent networks. The meaning of "softmax output" is explained in the main body of the text.

Figure 3. The feed-forward network classifies the C token in "Sales and Marketing"



Source: OECD elaboration on Devlin et al. (2018[34]).

Optimising the algorithm: fine-tuning BERT

57. Optimising the assignment of a category to each skill, as explained in the previous section, relies on the possibility for the algorithm to "learn" which of the skill-category associations it produces are incorrect and should therefore be improved upon in the following iteration of the algorithm. In a supervised setting, this is obtained by providing the algorithm with a subset of skill-category combinations that are assumed to be correct because they are produced by the researchers, who manually classify a subset of Burning Glass skills into their respective category. This is the so-called training set (see below for more details about its creation).

58. For each skill in the training set, the algorithm sees the pre-assigned, "true" category, and compares it to the output of the softmax prediction, i.e. the one predicted by the model. An "error", based on the difference between the predicted category and the "true" assigned category is computed. The algorithm's parameters are then chosen to minimise this error, following an iterative process. Through this process, the algorithm "learns" how to classify the Burning Glass skills, both those included in the training set and the others.

59. For each skill, the error that is minimised is a measure of dissimilarity, or cost function Z, between two vectors, p and q, where p is the "true" vector (which has value zero everywhere except at the index corresponding to the true category as assigned by the researcher, for which it has value 1) and q is the vector of predictions from the model (with values ranging from 0 to 1, with 58 possible values). Also called cross entropy, Z can be written as:

$$Z(p,q) = -\sum_{i=1}^{M} p_i \log(q_i)$$

60. This error measure is used to improve the features of the algorithm (parameters) via back-propagation: the error is used to update the last layer and this update is then used to compute new parameters for the second to last layer, etc. This process is repeated for the 12 layers of the model, and on each sub-sample of the training set, until the model's performance no longer improves. This process is called **fine-tuning** the model for the classification task.

About the softmax output

61. As mentioned, the final output of the optimisation process is a positive vector whose values sum to 1. In what follows, the vector's maximum value, which defines the predicted category, is called q, i.e. $q = \max(q_i)$ for each skill s, where (q_1 ... q_{58}) is the vector of predictions from the model as in the previous paragraph. In Figure 3, for instance, $q = 0.86$.

62. This vector (including q), however, cannot be interpreted as a set of unbiased probabilities assigning a skill to one of the 58 ONET+ categories. While predictions derived from multinomial logistic regressions or shallow neural networks are representative of the true underlying probability distribution with which items get assigned to a category, recent work has shown that this is no longer the case for modern neural networks (Guo et al., 2017[49]).[15]

63. In light of such bias, an observed $q = 0.86$ cannot be interpreted as the skill having an 86% chance of being correctly classified. For the same reason, the values of q of two distinct skills cannot be compared. Furthermore, while different values of q_i for the same skill can be ranked, and indeed max(q_i) = q, the actual numerical distance across q_i's for the same skill cannot be interpreted, and the vector output cannot be used as a set of weights for the importance of the 58 categories for the given skill. In other terms, the information contained in q_i is purely ordinal.

64. That said, q still conveys important information about the probability of misclassification or the confidence with which the algorithm classifies the keywords, since it is correlated with the model's accuracy. This notion will be developed in Section 5.

Training BERT for the purpose of this study

Finding definitions

65. Relying on the definition of a skill for its classification requires a definition for each of the 17 000 skills in Burning Glass. The model does not impose constraints to the nature or length of the definition, but its likelihood to classify the skill unambiguously increases with the informative power of the definition.

66. The first source of definitions is the European Skills/Competences, qualifications and Occupations (ESCO) database, which includes 13 485 skills and their definition. Only a few of Burning Glass skills, however, have the same designation in Burning Glass and ESCO, allowing for a perfect match.

[15] The source and potential solving of this type of bias in the predicted probabilities is still an open field of research (Sensoy, Kaplan and Kandemir, 2018[57]; Możejko, Susik and Karczewski, 2018[58]). The prevalent explanation offered in the literature is the lack of a default category to represent uncertainty: if the model encounters a skill that does not fit in any of the predetermined categories, it has no options to discard or flag the observation, as a prediction is wanted for all skills. The model is thus forced to produce a value for each of the predetermined categories.

Approximate matching[16] extends the overlap between the two databases. In total, an ESCO definition is attributed to 674 Burning Glass skills.

67. The remaining definitions are found by scraping Wikipedia. A Python module that wraps the MediaWiki API retrieves the summary of any Wikipedia article. The module is flexible enough to search for variations of skill keywords. For instance, requesting "Business Planning" will redirect to the "Business Plan" article. The retrieved definition consists of the first lines of the article. Wikipedia is a widely used corpus for research in computational linguistics, information retrieval and natural language processing (Medelyan et al., 2009[50]). Its ever-growing size, structure, wide coverage, and high quality constitute important advantages over other corpuses, and attract many researchers in these areas (Yano and Kang, 2016[51]). Another interesting feature of Wikipedia is that it covers a wide variety of scientific topics, in most cases with a high degree of sophistication (Thompson and Hanley, 2017[52]). This is particularly relevant in the context of this work since many keywords from Burning Glass are highly specialised terms (e.g. "Veritas NetBackup", "ARP4754").

68. The combined approaches produce a definition for 99% of the Burning Glass skills. The remaining 180 skills (accounting for 0.03% of all Burning Glass skills) cannot be found in either ESCO or Wikipedia, often because they are too specific or refer to new fields or technologies (e.g. "AdvantX Software", "PyBrain", "Roostify"). Consequently, these skills are not classified.

The training set

69. The optimisation of the model requires a labelled dataset, referred to as training set. A training set is a collection of examples of skills and their definition with their correct label, i.e. the category in which they should be classified. The training set indicates to the algorithm how the information in the embedded definition maps onto one of the O*NET+ categories. The algorithm therefore leverages at the same time knowledge from BERT and from the training set.

70. The quality of the training set has a first-order impact on the performance of the classification algorithm. A first dimension of interest is the size of the training set. A longer list of skills with their matched correct label has the potential to extend the information set the algorithm can leverage in the optimisation exercise. That said, no "ideal" size of the training set exists, and increasing the set's size fast hits a time and resource constraint. Furthermore, adding classified skills may fail to provide additional information, for instance when one adds a synonym of a skill already included in the training set, or a closely related concept.

71. Two other features of the training set are of primary importance: *homogeneity* and *universality*. Homogeneity requires that approximately the same number of skill keywords be included in each category. Without homogeneity, the model may have insufficient knowledge about some categories, and could fail to identify the corresponding skills or be tempted to prioritise categories that are overpopulated with skills in the training set. Universality requires that the training set include representative content in each category. Without universality, the model may miss some skills that are part of a category: for instance, having no film-related skills in the "Fine Arts" category of the training set will prevent the model to identify film-related Burning Glass skills as "Fine Arts".

72. In the present study, the training set is composed of 500 skills, a reasonable number in light of resource constraints. They were chosen as:

- the 200 most frequent skills in Burning Glass, and

[16] To identify suitable approximate matches, first a vector is associated to each keyword using a TF-IDF vectoriser. Approximate matches are then defined as pairs for which the cosine similarity between the two vectors is higher than 0.8.

- 100 randomly sampled skills in Burning Glass,[17] and
- 200 skills not appearing in Burning Glass but specifically chosen by the researchers.

73. The latter have been selected by the authors of this study based on their knowledge and understanding of the different O*NET+ categories, in order to improve coverage of the different aspects of a category and therefore to ensure universality. Choosing the skills also helps improving homogeneity: categories whose skills are rare by default in Burning Glass would be otherwise given lower priority in the classification exercise, in virtue of the limited knowledge the algorithm would have of them. For categories disproportionally including skills that are rare in Burning Glass, a random sampling of Burning Glass would indeed fail to provide a sufficient amount of skills to complete the training set. Adding 200 out-of-sample skills ensures that each categories contain at least three skills in the training set.

74. These 500 skills are classified manually by the authors into one of the 58 ONET+ categories. To bolster the quality of the training set, difficult cases are consistently cross-validated and discussed internally when necessary. Furthermore, randomly chosen pairs of skills with their attributed category are closely verified. Lastly, skills receiving two different categories from different researchers are thoroughly examined and disambiguated by the team, by cross-referencing each other's answers to achieve consensus. This happens in 25% of the cases, meaning than the degree of agreement between researchers during this manual allocation exercise is equal to 75%.

Implementation, training and choice of the hyper-parameters

75. As mentioned, the algorithm learns from the training set, and each skill of the training set must be shown at least once to the model. As the model is unable to process all 500 skills at once for computational reasons, BERT splits the training set into subsets. Instead of processing them one by one, subsets are grouped in batches of fixed number of skills (the **batch size**). This improves and accelerates the algorithm's learning. An **epoch** is counted each time all batches have been processed by the model. Usually, neural networks are trained on several epochs. The extent to which the parameters of the model change with each update depends on the **learning rate**. The batch size, the number of epochs and the learning rate are the so-called "hyper-parameters" that must be chosen before the training begins, and can depend in an important way by the availability of computing power.[18] While they influence the final performance of the model, the optimal value for each of them always depends on the problem at hand. To select the best possible values, the performance of the algorithm is benchmarked under different combinations of hyper-parameter values.

76. The first definition of performance is given by the **loss**, i.e. the sum of the cost function (the cross-entropy defined in the previous section) over all training samples. A lower loss means that the model has learned to minimise the distance between its output distribution and the "true" distribution provided by the training set. Another definition of performance can be given by the **accuracy**, which is the proportion of skills that the model correctly allocates to the appropriate category. Other indicators of performance and a more in-depth reflection on the relevance of accuracy are proposed in Annex D.

77. Deep neural networks may have a tendency to learn by heart the training data and loose generalisation capacity (a problem known as **over-fitting**). Thus, the performance cannot be measured on the training data, as it would overestimate the performance on unseen data. A **test set**, which includes 400 skills randomly selected from Burning Glass, which are labelled but not leveraged during the algorithm's training, can be used to compute an unbiased assessment of accuracy. Accuracy is calculated by asking the trained model to predict the category of the sub-set of keywords for which the classification

[17] Sampled with weights proportional to their frequency in Burning Glass data.

[18] Other hyper-parameters exist besides batch size, number of epochs and learning rate, but they have marginal impact on the results and are therefore neglected in this discussion.

is already known, i.e. the test set, and calculating the frequency of misclassified keywords. In the present study, the performance of the model is relatively stable, i.e. it does not dramatically change for different values of the hyper-parameters. The specific values of the hyper-parameters that maximise accuracy are therefore chosen to produce the final output of this study.[19]

[19] The specific values are the following: epoch=30; batch size=8; learning rate=2^(-5); maximum number of tokens for a definition=256.

5 The classification

78. The approach described so far yields a classification of each Burning Glass skill keyword into one ONET+ category. While reporting the full classification in a document is impossible, this section describes the results of the taxonomy exercise and presents descriptive statistics of the model's accuracy. Section 6 further describes the outcome of the classification by describing skills demand in the Burning Glass dataset, once skills are grouped according to the proposed classification.

A brief description of the classification output

79. The final taxonomy contains 17 331 skills grouped into 61 categories. There are 289 skills per category on average, while the median number of skills per category is equal to 108.

80. Figure 4 shows the absolute number of keywords contained in each O*NET+ category. The most populated categories are "Medicine and Dentistry", "Management of Financial Resources", and "Biology**",** including respectively 3 895, 1 070, and 1 044 keywords. The disproportionate content of the "Medicine and Dentistry" category stems from the high frequency of health-related job postings in the Burning Glass database (US SOC occupations 29 and 31 represent 12.4% of the total number of job postings in the pooled dataset[20]), as well as from the medical discipline's propensity to include a vast range of specialisation and tools, which translate in many unique or very specific skills in the Burning Glass dataset (e.g. specific medical conditions). The figure, however, does not suggest that skills in Medicine & Dentistry are in highest demand in online job postings, since it does not factor in the frequency of these keywords in job postings.

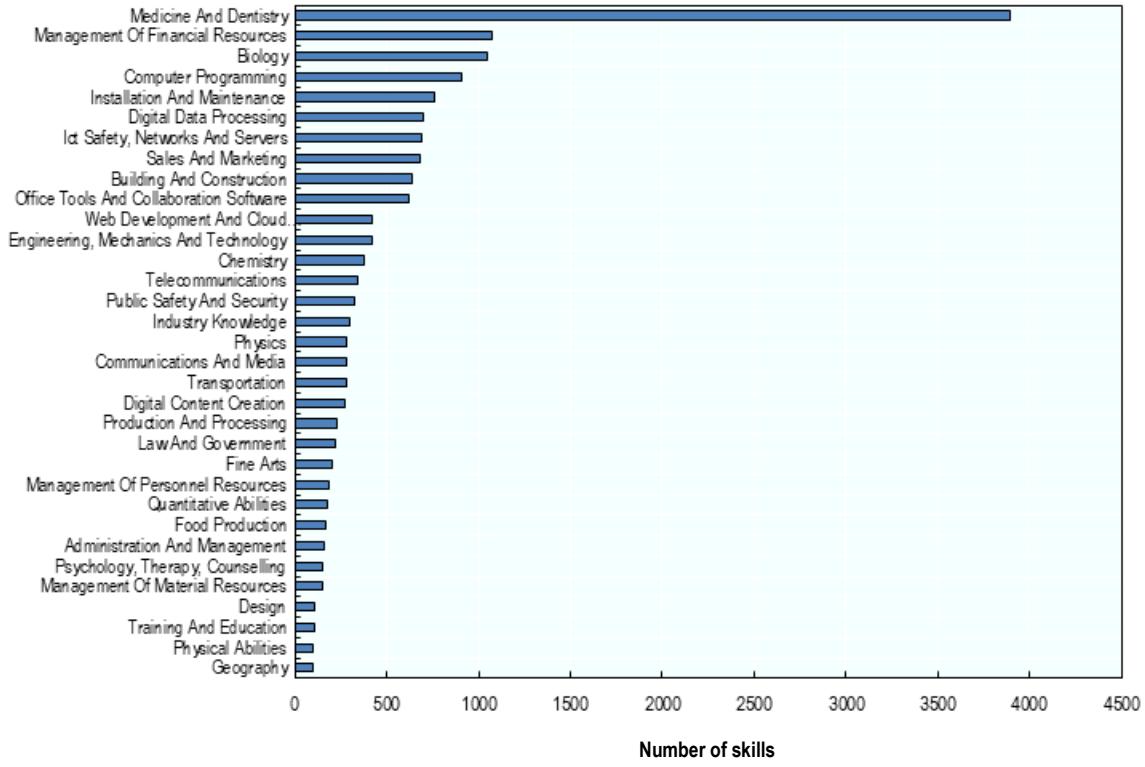
81. On the other hand, skills relating to human values or personality traits (e.g. falling into categories of "originality", "motivation and commitment", "adaptability and resilience", "self-management and rigour") are very rare in the list of skills contained in the Burning Glass data.[21] This may be the case because these competences are implicit and not clearly stated in the job advertisement, because employers do not use several synonyms to express the same concept, or because they are mentioned in the text of the job advertisement, but are not extracted by Burning Glass. The possibility that the algorithm is unable to classify appropriately the keywords related to values and personality traits cannot be ruled out completely, but it is not clear why the algorithm would perform worse for these skills than for others.

[20] These figures are calculated on the pooled dataset across countries and years. As a benchmark, in the United States, these two occupations represented 8.4% of the employed population (16 year olds and older) in 2018.

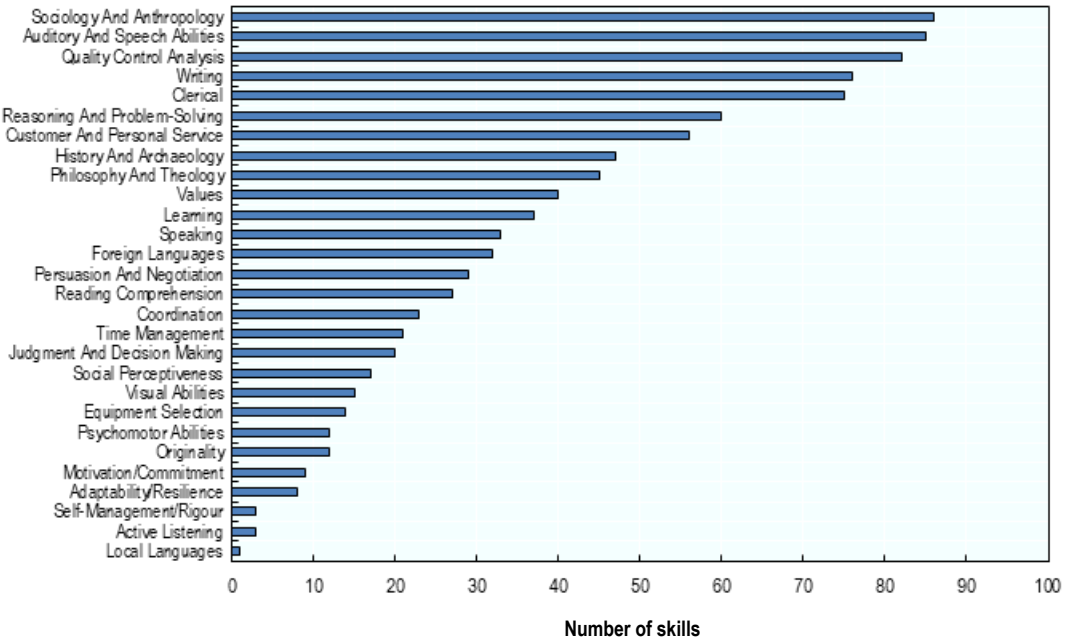
[21] This observation is based both on the fact that categories for attitudes and values are among the least populated in the taxonomy (see Figure 4), and on a visual inspection of Burning Glass skill keywords.

Figure 4. Number of skills per O*NET+ category

Panel A – Most populated categories (containing more than 100 keywords)



Panel B – Least populated categories (containing less than 100 keywords)



Notes: Pooled dataset: AUS, CAN, GBR, NZL and USA data, for the period 2012-2018.
Source: OECD calculations based on Burning Glass data.

Quantitative evaluation of the results

Model accuracy

82. As mentioned, accuracy is the proportion of skills that the model correctly allocates to the appropriate category. In the present context, it is the most important available indicator of the algorithm's performance. The minimum accuracy measured on the test set with the optimal configuration of the model is 74%, but this rises to 85% when excluding from the sample the least frequent skills (appearing in less than 10% of job postings).

83. Evaluating whether this degree of accuracy is satisfactory is not straightforward. Comparisons with the literature are seldom meaningful, as classification exercises can differ very much in nature, and often require a simpler classification of items into two categories only. For instance, the standard General Language Understanding Evaluation (Wang et al., 2018[53]) performs several tasks, but many are binary, such as assessing whether a sentence is an entailment, contradiction, or neutral with respect to another, or whether two questions are semantically equivalent. For these simpler classification tasks, BERT achieves accuracies ranging from 65 to 95%. The present classification exercise therefore achieves accuracy in line with other, simpler, real-world applications of the BERT model.

84. The realised accuracy can also be compared with the hypothetical accuracy that naïve classification models would achieve, and with human accuracy in a manual classification. Customarily, two naïve models are proposed and used as benchmark: the random rule and the zero rule. Following the random rule approach, each skill is attributed to a category at random following a uniform distribution. The resulting accuracy is logically $1/61 \approx 2\%$.[22] Under the zero rule, the most populated category (as estimated on the training set, here "Medicine and Dentistry") is attributed to all skills. This approach achieves a 14% accuracy. Finally, several experiments conducted during the validation of the training set indicate that humans agree only 75% of the time approximately. Additional evaluation metrics are provided in Annex D.

85. Several reasons could be advanced to explain why some skills are not attributed to the correct category. First, many Burning Glass skills are vaguely phrased or incongruous (e.g. "Bowling", "Human Guides", "VTPSUHM7", "HORVIP", etc.). Even an expert could face difficulties to assign them to any category. Second, definitions of skill keywords may not be sufficiently long or precise for the model to understand fully the skills' meaning. Note that this last point is not a property of rare skills. The process of acquiring a definition is independent from the frequency of the skill in Burning Glass. Thus, rare skills can still be well defined and well classified, as in the case, for instance, of skills falling in the "attitudes" category, and the methodology can be considered robust to the rarity of skills.

Relationship between accuracy and q

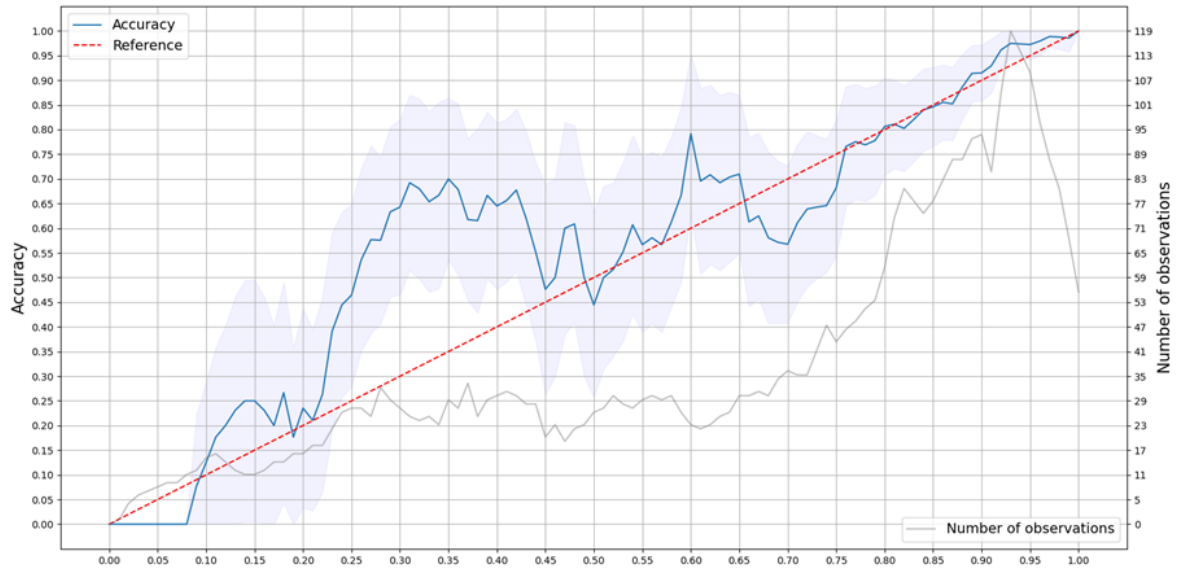
86. As previously explained, the value q is the maximum value of the vector obtained after the optimisation process, and it defines the predicted category. While it cannot be interpreted as a probability, it is correlated with the model's accuracy and hence can give an approximate idea of the confidence one can have in the quality of the classification output.

87. Figure 5 shows the accuracy of the model as a function of q (blue curve). Based on the final specification of the hyperparameters, for any $\hat{q}$ in [0,1], the test set is restrained to the skills which have a q in $[\hat{q} - 0.05, \hat{q} + 0.05]$, and the accuracy is computed on this subset of the test set. The blue shade is the

[22] A less naive random rule would attribute skills to different categories using a weighted distribution estimated from the training set (based on the number of skills per category). This less naïve approach yields accuracy ~4%.

95% confidence interval for the blue curve[23], and the grey curve represents the number of skills used to compute the accuracy. The 45-degree line is dashed and represents the locus of skills for which q and accuracy correspond.

Figure 5. Accuracy as a function of q



Note: x-axis reports values of q.
Source: OECD calculations based on the test sample constructed from Burning Glass data.

88. As mentioned in Section 4, accuracy is not equal to q, and indeed blue and dashed lines are not perfectly overlapping. However, accuracy and q are positively correlated. This means that the model tends to be more accurate when it produces a high q, and that filtering out skills with a low value of q can increase the global accuracy on the remaining skills classification. Table 4 provides the minimum value of q needed to reach a given level of accuracy (as computed on the test set) and the number of skills remaining in the Burning Glass dataset when choosing such value of q. For example, an accuracy superior to 85% can be obtained by excluding the skills with $q < 0.55$. By doing so, 11 222 skills (91% of observations in the skill dataset) remain in the dataset.

89. The distribution of q over all classified Burning Glass skills, shown in Figure E.1, is actually highly skewed to the left, meaning that the q value for a significant share of classified skills is close to 1. Excluding skills with low q values therefore does not reduce significantly the number of skills to work with.

Table 4. Accuracy as a function of q

Desired accuracy	74%	80%		85%	90%	95%
Threshold for *q*	0.00	0.35		0.55	0.74	0.85
% of postings included	100	95		91	87	82

Source: OECD calculations based on Burning Glass data.

[23] This confidence interval is computed assuming that each skill classification is a Bernoulli trial (correct/incorrect) with constant probability p. However, this assumption is likely violated, as the model is probably better at classifying some skills (e.g. Medicine) than others (e.g. Values). Therefore the values for the confidence interval should be considered suggestive.

Accuracy for broader categories

90. Importantly, the predicted category is not necessarily far from the "true" category even when the model is not correct. As the O*NET+ taxonomy contains categories that are close in concept ("Computer Programming" and "Digital Data Processing", for example), the algorithm may find it difficult to distinguish between two similar categories, especially when the skills' definitions are short or ambiguous. Under a more aggregated version of the taxonomy, these categories may be merged into a single one, which the algorithm could now correctly identify as the relevant category for a given skill, even when the prediction was incorrect at the more disaggregated taxonomy level. This is relevant because 61 categories may still be too many in an empirical analysis, so that a more aggregated version of the taxonomy may be useful. Annex B describes a hierarchical mapping of the 61 ONET+ categories into 16 more aggregate categories.

91. Using this taxonomy of broader categories, the algorithm achieves an accuracy of 82%. Since accuracy increases mechanically when considering fewer categories, the benchmark values need adjusting as well (Table 5).

Table 5. Achieved accuracy and benchmark values for two level of aggregation of the final taxonomy

	O*NET+ with 61 categories	Higher level taxonomy with 16 categories
Model accuracy	74%	82%
Random rule accuracy	2%	6%
Zero rule accuracy	14%	23%

Source: OECD calculations based on the test sample constructed from Burning Glass data.

6 Further validation

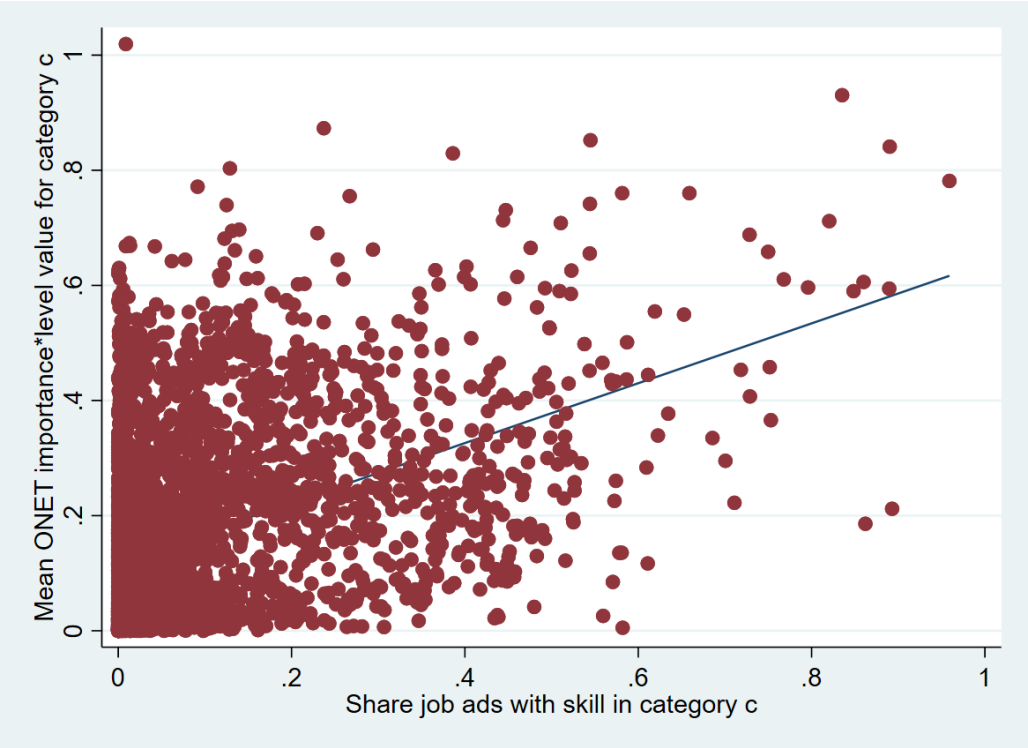
92. The validity of the classification can also be assessed indirectly, leveraging other sources of data and existing results from the literature. In this section, a first exercise compares the outcome of the classification of Burning Glass skills with data reported in O*NET. A second set of exercises proposes some descriptive statistics using the reclassified data, and compares them to the results obtained by the literature, with the goal of assessing their coherence. For simplicity, evidence in this section is generated from data for Canada in 2018, but similar analyses can be conducted for other countries and years.

Comparison with O*NET

93. This subsection compares the outcome of the classification of Burning Glass skills and data on skill requirements as reported in O*NET. Figure 6 correlates the share of job postings requiring at least one skill keyword in category *c* with the average importance*level value for O*NET category *c,* by occupation (U.S. SOC 4-digits). The association is not expected to be perfect, even if all skill keywords were perfectly classified. First, O*NET reflects the skill content of occupations in the entire economy, while Burning Glass data capture a snapshot of the *online vacancy* market, which is not expected to perfectly represent the entire labour market. Second, skill measures by occupation in O*NET and in Burning Glass are conceptually different: O*NET reports the level of a skill needed to perform one occupation and the importance of this skill for the occupation, while Burning Glass reports which skills are requested in which job advertisement, information which is here summarised as the share of job postings in one occupation that require a given skill.

94. These caveats notwithstanding, the figure shows a positive and significant relationship between the two indicators, and one that also holds when looking at U.S. SOC 2- and 3-digit disaggregation levels, or when including 2-digit occupation fixed effects, which indicates that the correlation is not driven by one particular occupational group.

Figure 6. Correlation between the classification results and O*NET importance values, by occupation



Note: One point corresponds to a skill category for one occupation (U.S. SOC 4-digits). Y-axis: importance*level value (rescaled to fall between 0 and 1) for O*NET category *c* by occupation, averaged across all underlying 6-digit U.S. SOC occupations in the 4-digit occupation. X-axis: share of job ads requiring at least one skill in category *c*, by 4-digit occupation. Job postings with more than 20 skills and occupations with less than 200 observations are excluded to limit noise in the estimation. The blue line represents the linear prediction resulting from the regression of O*NET importance*level value of category *c* onto the share of job ads requiring at least one skill in category *c*. The R^2 of this regression is equal to 0.19. The correlation between the two variables is equal to 0.52. Only categories that are exactly similar in O*NET and the classification are kept here (additional digital skills are thus absent, and merged categories are removed).
Source: OECD calculations based on Burning Glass data for CAN (2018).

Comparison with results from the literature

95. As mentioned, a different, indirect way to validate the classification in this study is to replicate some of the evidence from the existing economic literature relying on Burning Glass or other data. This subsection presents three such examples that describe, respectively: the proportion of the variation in the skill requirements that can be attributed to different characteristics of the job, as reported in the advertisement; the different probability with which a given skill is required in jobs opened in digital intensive vs. less digital intensive sectors; and the wage premium associated to a given skill requirement, holding other features of the advertised jobs constant.

Variance in skill requirements

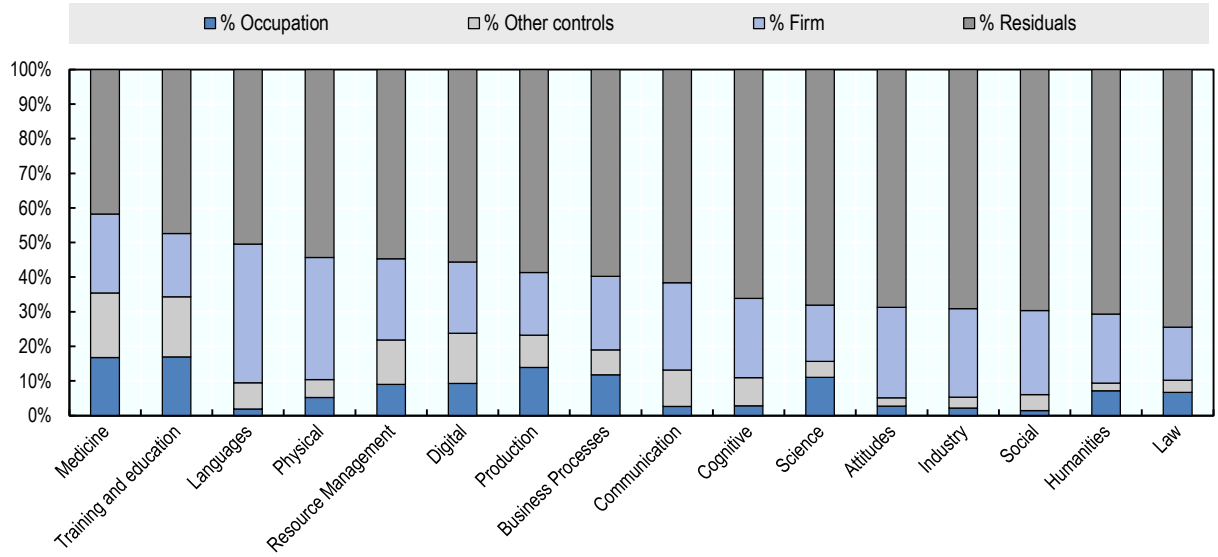
96. One possibility to infer (online residual) demand for certain skill categories is to compute the share of job postings listing at least one keyword belonging to those categories, which gives the empirical probability of job postings requiring a given skill. This is the approach taken by an important part of the literature (see for instance Deming and Kahn, 2018[1]). A first exercise explores how the observed variation

in skill requirements across job postings is related to variation in observed (occupation, firm, location, education and experience requirements) and unobserved dimensions of the advertised job.

97. Figure 7 shows the results of a regression, at the job posting level, of a dummy variable indicating whether the advertisement lists at least one skill in category *c,* onto education and experience requirements and 6-digits occupation, region, and firm fixed effects. The exercise is repeated for the sixteen categories of the higher-level hierarchy. As in Deming and Kahn (2018[1]), to limit noise in the estimation, the sample is restricted to firms with at least two ads in 2018, to professional occupations (SOC codes 11 to 29)[24], and to job postings listing less than 20 skill keywords. The height of each bar represents the total variance for each category. This overall variance in the requirement of a particular skill category is decomposed into the variance in the fitted values for the occupation fixed effects (bottom bar in dark blue), for job location and education and experience requirements (light grey bar), for the firm fixed effects (light blue bar), and into the remaining unexplained variance (dark grey bar).[25]

Figure 7. Sources of variance in skill requirements across ads

Variance by underlying factor



Note: We regress an indicator taking value 1 if the job advertisement contains at least one skill in the specified category, and 0 otherwise, on a set of controls including: an indicator for the 6-digit U.S. SOC 2010 occupation ("occupation"), an indicator for the employer ("firm"), and "other controls" (education and location dummies, and minimum required experience). 100% corresponds to the variance of the residuals and of the fitted values based on specified controls. The sample is restricted to firms with at least two ads in 2018, to professional occupations (U.S. SOC codes 11 to 29), and to job postings listing less than 20 skills.
Source: OECD calculations based on Burning Glass data for CAN (2018).

98. Occupation fixed effects account for a small fraction of the total variance in skill requirements (less than 10%). This percentage is slightly higher for Business Process, Production, Medicine, Science, and Training and Education categories. Controls for job location and education and experience requirements ("Other controls") also account for less than 10% of the total variation, with the exception of Medicine, Management, Digital, and Training and Education categories. On the contrary, there is substantial variation across firms in their tendency to specify the different skill requirements: between 15% and 40% of total

[24] Results are qualitatively similar when including all occupations (U.S. SOC codes 11 to 53).

[25] Sector fixed effects are not included because they would be absorbed by firm fixed effects in most cases.

variance is explained by firm fixed effects, with Languages and Physical skills being the two categories for which the identity of firm is the most important. These can be explained by systematic differences in firms' recruiting strategies and in skill utilisation. However, most of the variance in each skill requirement (between 50% and 70%, depending on the skill category) remains unexplained with this set of control variables and firm fixed effects.[26] These results are consistent with findings in Deming and Kahn (2018[1]) for the United States for the period 2010-2015, although the unexplained variance is lower in their case (about 50% on average). Among skill categories with the highest total variance, cognitive skills are those for which the largest part remains unexplained. Other categories for which the unexplained component is substantial are social skills, attitudes, and several knowledge categories (science, law, humanities and industry-specific knowledge).

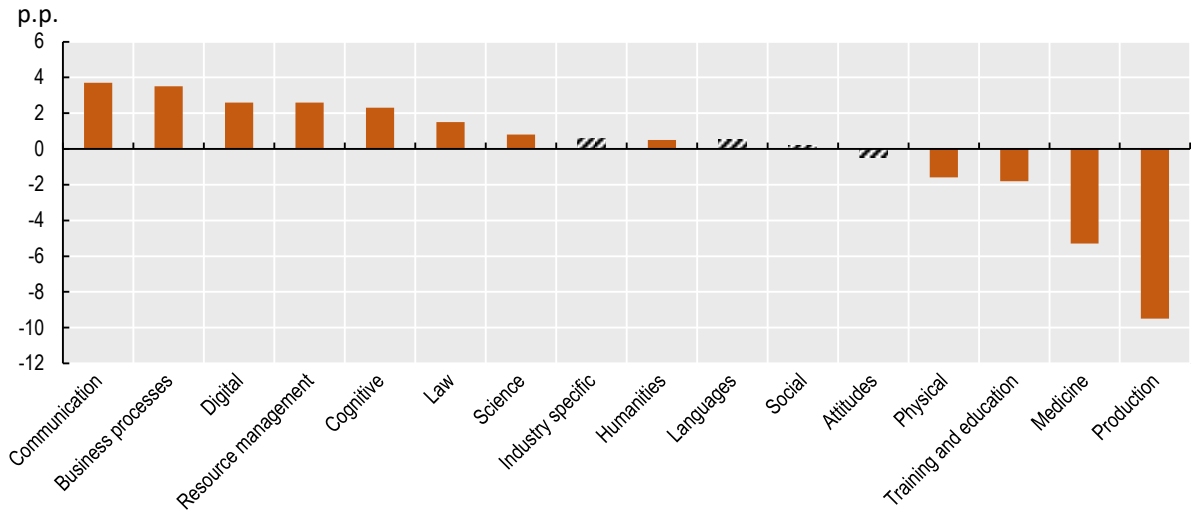
Sectoral differences in skill requirements

99. This second exercise exploits the variation in skill requirements related to the employer's sectoral affiliation. In particular, it distinguishes between sectors that are more vs less intensive in digital technologies, as classified by Calvino et al. (2018[54]). Sectors can then be classified as digital intensive if they lie above the median of digital intensity across sectors, or less digital intensive if otherwise. Indeed, digital intensive sectors see a more pervasive diffusion of digital technologies, as captured by seven different intensity dimensions (ICT and software investment, purchases of ICT intermediate goods and services, use of robots, employment of ICT specialists; revenues of online sales), and this is likely to have a significant impact on skill need. The analysis above shows that firms' characteristics explain an important part of the observed differences in skill requirements between job postings. One particular characteristic that could drive the differences is the sector in which a firm operates, and more particularly its digital intensity. The link between digital-intensity and skill requirements has been explored empirically by Grundke et al. (2018[3]), but they do not look at skill requirements in online postings (they use data on skill use on the job from the OECD Survey of Adult Skills) and rely on a more limited number of skills.

100. Figure 8 shows how skill requirements differ in online postings for jobs in digital intensive *vs* less digital intensive sectors, after factoring-in a number of other features of the job postings such as location, occupation and employer. Even for the same employer, jobs in digital intensive sectors require more frequently skills related to communication and digital technologies themselves, but also managerial skills ("Business processes", "Resource management") and higher cognitive abilities ("Cognitive skills", "Scientific knowledge"). Conversely, physical skills and skills related to production – among others – are more frequently demanded in postings for jobs in less digital intensive sectors. By design of the estimation, these results are not necessarily driven by the fact that jobs in digital intensive sectors are also more intensive in occupations related to ICT.

[26] The residuals capture factors affecting skill requirements in job advertisement, but which are not linked to employer, sector, occupation, geographical location, education and experience requirements. One such example is differences across job titles, i.e. more detailed occupational specifications.

Figure 8 Difference in the probability of requiring a certain skill, digital intensive vs less digital intensive sectors



Note: The graph plots the marginal effect of an OLS regression, where the probability that a job advertisement requires at least one skill in a given skill category is regressed on an indicator variable with value 1 if the job is advertised in a digital intensive sectors according to Calvino et al. (2018[54]) and zero otherwise, and dummies for State, 2-digit U.S. SOC 2010 occupation and employer name. Shaded bars identify differences (coefficients) which are not statistically significant at the 5% confidence level based on robust standard errors.
Source: OECD estimations based on Burning Glass data for CAN (2018).

Heterogeneity in wage returns to skill requirements

101. A last application of the classified database assesses whether certain skill categories are associated with a higher wage premium.[27] Figure 9 reports the estimated correlation between the hourly wage posted in a job advertisement, and the probability that the advertisement requires at least one skill in a given skill category. All skill categories are pooled, as opposed to estimating the wage returns to each category separately, to avoid that the coefficient on any single category captures the cross-correlations among skills. The estimation further controls for required experience and educational attainment, and for dummies for the State, sector and occupation reported in the job advertisement.

102. Once one considers all skills at the same time, a positive premium is associated with industry-specific skills, but also cognitive, science, social and especially management and legal skills, irrespective of sector, occupation or geographical location of the advertised job. Conversely, postings requiring physical skills, skills most related to production or business processes, language skills and attitudes offer on average lower starting wages. The relationship to hourly wages is not statistically significant for the remaining skill categories.

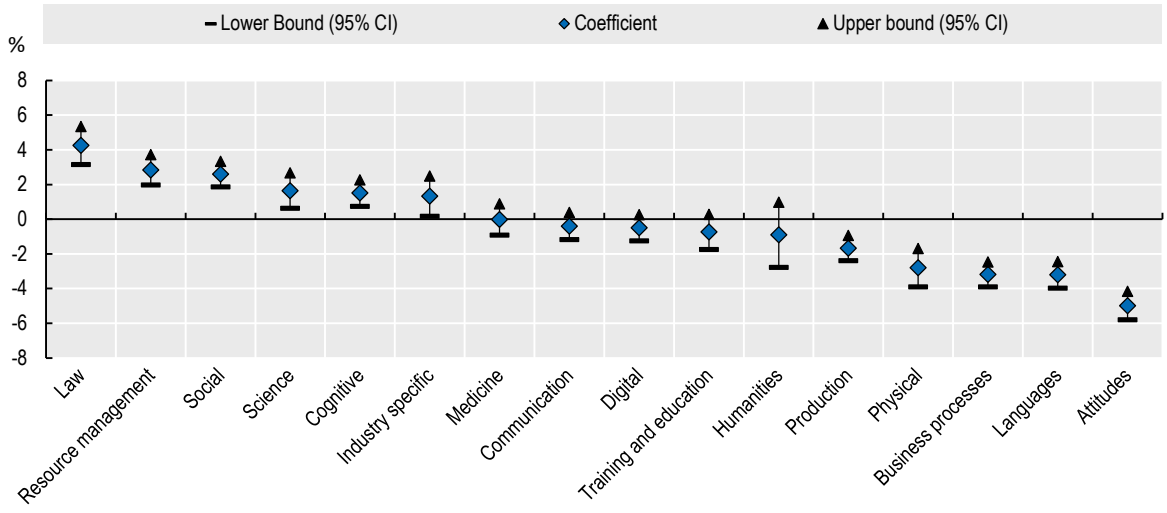
103. These results are similar to findings by Deming and Kahn (2018[1]) who, using Burning Glass data for the United States, show that social and cognitive skill requirements in job ads are positively correlated with occupational wage differences across local labour markets, even after controlling for education and experience requirements, other skill requirements, location fixed effects, and industry and occupation (6-digit) fixed effects. They also find that the premium associated to social skills is much larger than the one

[27] These are wages posted in the job advertisement and, as such, typically exclude benefits and bonuses and are likely subject to renegotiation during the hiring process. In addition, only a subset of ads in the Burning Glass dataset include wage information and they are likely to be a selected sample. For Canada in 2018, this proportion is about 70%.

for cognitive skills. Moreover, the findings on cognitive skills in the present paper are coherent with wage returns estimated on administrative data: using the OECD Survey of Adult Skills (PIAAC): Hanushek et al. (2015[4]) estimate labour market returns to cognitive skills of 19 percent in Canada, as opposed to a 15 percent return in the present paper. The positive premium found for science skills is also consistent with previous research using Burning Glass data for the United Kingdom which shows that STEM jobs (i.e. those with STEM skill requirements) are associated with higher wages than non-STEM jobs (Grinis, 2019[5]). Results on digital skills warrant more consideration. Indeed, digital skills fall in the group of STEM skills, but are found to have no significant wage returns in our exercise (Figure 9). This may reflect a number of factors. First, the same digital competence may have a substantial wage premium in some occupations, sectors or geographies, but not in others, a fact that the coefficient on "Digital" in Figure 9 cannot capture, as the variation is controlled for by other terms in the estimation. Second, digital skills may not be especially rewarded in isolation, but only when combined with other, complementary skills, something which is accounted for by other factors in the present estimation. Lastly, skills classified as "Digital" are heterogeneous in nature in the Burning Glass dataset, ranging from basic to very sophisticated capabilities, which can therefore have substantially different wage premia. This hypothesis was tested in Figure F.1 in Annex F, which reproduces the exact same specification as Figure 9, but substituting the aggregate category of "Digital" skills with its subcomponents, as per Table B.1. Indeed skill requirements in the different digital subcategories are associated to wage premia of different magnitude and sign. In particular, everything else held constant, higher posted wages are associated to computer programming skills and skill related to ICT safety or the construction of ICT networks only. Wages in postings requiring the other "Digital" skills are lower, likely because some such skills are less sophisticated in nature.

Figure 9 Hourly wage elasticity by skill requirement

Percentage change in hourly wages if the job posting requires at least one skill in the category, everything else held constant.



Note: The graph plots the coefficients of a single estimation on Canadian 2018 job postings data, where the logarithm of posted hourly wages is regressed on required experience, dummies for educational attainment, dummies for all skill requirements (one for each of the 16 "Broad" categories presented above), State, 2-digit ISIC rev.4 industry and 2-digit U.S. SOC 2010 occupational dummies. Dummies for skill requirements take value 1 if the job posting requires at least one skill falling into the category on the x-axis.
Source: OECD estimations based on Burning Glass data for CAN (2018).

7 Conclusions

104. The present study introduces a semi-supervised machine learning approach to the classification of the approximately 17 000 skill keywords appearing in the Burning Glass dataset at the moment of writing. The approach relies on three main inputs: (i) a taxonomy of 61 skill categories to which the skill keywords should be related; (ii) a training set, i.e. a collection of examples of skills and their definition with their correct label or category, as manually classified by the authors according to the skills' meaning; and (iii) the definitions of the skill keywords, as derived from ESCO and Wikipedia. By relying on the meaning of the skill keywords, the approach accounts for duplicates and synonyms in the Burning Glass datasets, and produces a classification that is stable over time.

105. The final taxonomy, or list of 61 mutually exclusive skill categories, largely builds on O*NET with extra categories for digital skills. The O*NET database is developed and validated by labour market and education experts that provides detailed information on skill requirements in occupations and that organises this information under a clear hierarchical structure.

106. The resulting taxonomy is then used to construct the training dataset, i.e. a subset of Burning Glass skills that the authors manually classify in the relevant ONET+ skill category according to their meaning. In this study, the set contains approximately 500 skill keywords chosen to ensure its homogeneity and universality.

107. Skills are classified into one of the ONET+ categories using BERT, a Natural Language Processing algorithm particularly suited for sentence classification. BERT associates a vector to each skill keyword based on the keyword's definition, thus allowing for the detection of semantic similarities across skill keywords. The algorithm is then trained to classify the skills keywords by iteratively comparing the results of its own classification with the manual classification contained in the training set.

108. The resulting classification has an estimated accuracy (the percentage of correct predictions of the model evaluated against a manually constructed test set) of 74%, which rises to 85% when restricting the output to skills that are less ambiguously classified by the algorithm and with minor loss of information. These figures are satisfactory, given the intrinsic difficulty to classify vague or ill-defined skill keywords, even with a manual procedure. For comparison, a manual classification of a random sub-set of skills achieved a 75% accuracy rate, while a random allocation of all keywords across the skill categories would obtain a 2% accuracy rate.

109. The last section of the study validates the results by comparing the outcome of the classification of Burning Glass skills with data reported in O*NET. A second exercise uses the skill classification derived in the paper to describe patterns of skill requirements in online postings, and compares them to the findings emanating from the existing literature. The results are largely consistent with the results found in the literature and with the O*NET data, providing an initial confirmation that the skills have been reclassified in a meaningful way.

References

Beblavý, M., B. Fabo and K. Lenearts (2016), *Demand for Digital Skills in the US Labour Market: The IT Skills Pyramid*, CEPS Working Paper. [11]

Blair, P. and D. Deming (2020), *Structural Increases in Skill Demand After the Great Recession*, NBER Working Paper. [13]

Börner, K. et al. (2018), "Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy", *Proceedings of the National Academy of Sciences*, Vol. 115/50, pp. 12630-12637. [9]

Bruening, N. and P. Mangeol (2020), *What skills do employers seek in graduates? Using Burning Glass Technologies job postings data to support policy and practice in higher education*, OECD Education Working Papers, n. 231, OECD Publishing, Paris, https://doi.org/10.1787/bf533d35-en. [21]

Burke, M. et al. (2019), *No Longer Qualified: Changes in the Supply and Demand for Skills within Occupations*, Working Papers 20-3. [14]

Burning Glass Technologies (2019), *Beyond Tech - The Rising Demand for IT Skills in Non-Tech Industries*. [6]

Burning Glass Technologies (2019), *No Longer Optional - Employer Demand for Digital Skills in the UK*. [8]

Burning Glass Technologies (2019), *What's Trending in Jobs and Skills*. [7]

Calvino, F. et al. (2018), *A taxonomy of digital intensive sectors*, OECD Science, Technology and Industry Working Papers, 2018/14, OECD Publishing, Paris, https://dx.doi.org/10.1787/f404736a-en. [54]

Cammeraat, E. and M. Squicciarini (2021), *Burning Glass Technologies' data use in policy-relevant analysis: An occupation level assessment*, OECD Science, Technology and Innovation Working Papers, 2021/05, OECD Publishing, https://doi.org/10.1787/cd75c3e7-en. [39]

Campello, M., J. Gao and Q. Xu (2019), *Personal Income Taxes and Labor Downskilling: Evidence from 27 Million Job Postings*, Kelley School of Business Research Paper No. 19-35. [18]

Carnevale, A., T. Jayasundera and D. Repnikov (2014), *Understanding online job ads data*. [38]

Cedefop (2013), "Quantifying skill needs in Europe. Occupational skills profiles: methodology and application", No. 30, Research Paper, n.30, Luxembourg: Publications Office of the European Union, http://dx.doi.org/10.2801/13390. [41]

Chung, J. et al. (2014), "Empirical evaluation of gated recurrent neural networks on sequence modeling", *arxiv:1412.3555*, http://arxiv.org/abs/1412.3555. [32]

Consoli, D. et al. (2019), *Routinization, Within-Occupation Task Changes and Long-Run Employment Dynamics*, SciencesPo OFCE Working Paper, 08/2019, SciencesPo, Paris. [45]

Dawson, N. et al. (2019), "Adaptively selecting occupations to detect skill shortages from online job ads", *arXiv:1911.02302*, https://arxiv.org/pdf/1911.02302.pdf. [23]

Deming, D. (2017), "The Growing Importance of Social Skills in the Labor Market*", *The Quarterly Journal of Economics*, Vol. 132/4, pp. 1593-1640, http://dx.doi.org/10.1093/qje/qjx022. [44]

Deming, D. and L. Kahn (2018), "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals", *Journal of Labor Economics*, Vol. 36/S1, pp. S337-S369, http://dx.doi.org/10.1086/694106. [1]

Deming, D. and K. Noray (2020), "Earnings Dynamics, Changing Job Skills, and STEM Careers*", *The Quarterly Journal of Economics*, http://dx.doi.org/10.1093/qje/qjaa021. [20]

Devlin, J. et al. (2018), *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv:1810.04805, https://arxiv.org/abs/1810.04805. [34]

Dillender, M. and E. Forsythe (2019), *Computerization of White Collar Jobs*, Upjohn Working Papers and Journal Articles 19-310, W.E. Upjohn Institute for Employment Research. [17]

Djumalieva, J. and C. Sleeman (2018), *An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts*, Economic Statistics Centre of Excellence (ESCoE) Discussion Papers. [22]

Dorrer, J. (2014), "Do We Have the Workforce Skills for Maine's Innovation Economy?", *Maine Policy Review*. [10]

Goodfellow, I., Y. Bengio and A. Courville (2016), *Deep Learning*, MIT Press, http://www.deeplearningbook.org. [29]

Goos, M., A. Manning and A. Salomons (2014), "Explaining Job Polarization: Routine-Biased Technological Change and Offshoring", *American Economic Review*, Vol. 104/8, pp. 2509-2526, http://dx.doi.org/10.1257/aer.104.8.2509. [46]

Gorodkin, J. (2004), "Comparing two K-category assignments by a K-category correlation coefficient", *Computational Biology and Chemistry*, Vol. 28/5-6, pp. 367-374, http://dx.doi.org/10.1016/j.compbiolchem.2004.09.006. [56]

Graves, A. and J. Schmidhuber (2005), "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", *Neural networks*, https://doi.org/10.1016/j.neunet.2005.06.042. [31]

Grinis, I. (2019), "The STEM requirements of "Non-STEM" jobs: Evidence from UK online vacancy postings", *Economics of Education Review*, Vol. 70, pp. 144-158, http://dx.doi.org/10.1016/j.econedurev.2019.02.005. [5]

Grundke, R. et al. (2018), *Which skills for the digital era?: Returns to skills analysis*, OECD Science, Technology and Industry Working Papers n.2018/09, OECD Publishing, Paris, https://doi.org/10.1787/9a9479b5-en. [3]

Guo, C. et al. (2017), "On Calibration of Modern Neural Networks", *arXiv:1706.04599*, https://arxiv.org/abs/1706.04599. [49]

Handel, M. (2012), *Trends in Job Skill Demands in OECD Countries*, OECD Social, Employment and Migration Working Papers N.143, OECD Publishing, https://dx.doi.org/10.1787/5k8zk8pcq6td-en. [43]

Hanushek, E. et al. (2015), "Returns to skills around the world: Evidence from PIAAC", *European Economic Review*, Vol. 73, pp. 103-130, http://dx.doi.org/10.1016/j.euroecorev.2014.10.006. [4]

Hershbein, B. and L. Kahn (2018), "Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings", *American Economic Review*, Vol. 108/7, pp. 1737-1772, http://dx.doi.org/10.1257/aer.20161570. [2]

Hochreiter, S. and J. Schmidhuber (1997), "Long short-term memory", *Neural computation*, https://doi.org/10.1162/neco.1997.9.8.1735. [30]

Howard, J. and S. Ruder (2018), "Universal language model fine-tuning for text classification", *arXiv:1801.06146*, https://arxiv.org/abs/1801.06146. [33]

Jones, K. (1972), "A statistical interpretation of term specificity and its application in retrieval.", *Journal of documentation*. [25]

Koucký, J., J. Kovařovic and M. Lepič (2012), *Occupational Skills Profiles: Methodology and application*, Charles University in Prague, Prague. [42]

Kowsari, K. et al. (2020), "Text Classification Algorithms: A Survey", *arXiv:1904.08067*, https://arxiv.org/abs/1904.08067. [24]

Kuhn, P., P. Luck and H. Mansour (2018), *Offshoring and Skills Demand*. [16]

Liu, Y. et al. (2019), *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, http://dx.doi.org/arXiv:1907.11692. [35]

Medelyan, O. et al. (2009), "Mining meaning from Wikipedia", *International Journal of Human-Computer Studies*, Vol. 67/9, pp. 716-754, http://dx.doi.org/10.1016/j.ijhcs.2009.05.004. [50]

Mikolov, T. et al. (2013), "Distributed Representations of Words and Phrases and their Compositionality", *Neural and Information Processing System*, http://dx.doi.org/doi:10.5555/2999792.2999959. [26]

Modestino, A., D. Shoag and J. Ballance (2019), "Upskilling: Do Employers Demand Greater Skill When Workers are Plentiful?", *The Review of Economics and Statistics*, pp. 1-46, http://dx.doi.org/10.1162/rest_a_00835. [12]

Modestino, A., D. Shoag and J. Ballance (2016), "Downskilling: changes in employer skill requirements over the business cycle", *Labour Economics*, Vol. 41, pp. 333-347, http://dx.doi.org/10.1016/j.labeco.2016.05.010. [15]

Możejko, M., M. Susik and R. Karczewski (2018), "Inhibited Softmax for Uncertainty Estimation in Neural Networks", *arXiv:1810.01861*, https://arxiv.org/abs/1810.01861. [58]

OECD (2017), *Getting Skills Right: Skills for Jobs Indicators*, Getting Skills Right, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264277878-en. [47]

Pennington, J., R. Socher and C. Manning (2014), *Glove: Global vectors for word representation*. [27]

Peters, M. et al. (2018), "Deep contextualized word representations", *arXiv:1802.05365*, https://arxiv.org/abs/1802.05365. [28]

Radford, A. et al. (2019), *Language models are unsupervised multitask learners.*, OpenAI Blog. [37]

Rothwell, J. (2014), *Still Searching: Job Vacancies and STEM Skills*, Brookings Institution. [19]

Schuster, M. and K. Nakajima (2012), *Japanese and Korean voice search*. [55]

Sensoy, M., L. Kaplan and M. Kandemir (2018), *Evidential deep learning to quantify classification uncertainty*. [57]

Thompson, N. and D. Hanley (2017), "Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial", *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.3039505. [52]

Tsacoumis, S. and S. Willison (2010), "O*NET Analyst Occupational Skill Ratings", *Human Resources Research Organization* FR-08-70. [40]

Vaswani, A. et al. (2017), "Attention Is All You Need", *arXiv:1706.03762*, https://arxiv.org/abs/1706.03762. [48]

Wang, A. et al. (2018), "Glue: A multi-task benchmark and analysis platform for natural language understanding", *arXiv:1804.07461*, https://arxiv.org/abs/1804.07461. [53]

Yang, Z. et al. (2019), "Xlnet: Generalized autoregressive pretraining for language understanding", *Advances in neural information processing systems*. [36]

Yano, T. and M. Kang (2016), *Taking advantage of wikipedia in natural language processing*. [51]

Annex A. Availability of skill information by occupation

Table A.1. Percentage of observations with non-missing skill requirements

3-dig US SOC Occupation	%		3-dig US SOC Occupation	%		3-dig US SOC Occupation	%
11-1	96.7		31-2	99.2		43-6	97.6
11-2	99.2		31-9	97.6		43-9	98.5
11-3	97.9		33-1	98.4		45-1	62.1
11-9	97.8		33-2	91.0		45-2	79.3
13-1	97.5		33-3	95.3		45-4	90.0
13-2	98.2		33-9	95.4		47-1	89.8
15-1	98.8		35-1	97.9		47-2	89.8
15-2	98.6		35-2	92.8		47-3	94.0
17-1	87.6		35-3	93.2		47-4	94.0
17-2	97.3		35-9	93.1		47-5	92.9
17-3	96.9		37-1	98.8		49-1	97.2
19-1	98.4		37-2	94.7		49-2	97.0
19-2	98.7		37-3	89.7		49-3	95.7
19-3	97.9		39-1	93.6		49-9	95.4
19-4	98.2		39-2	96.2		51-1	96.5
21-1	96.5		39-3	92.6		51-2	93.4
21-2	96.7		39-4	68.3		51-3	91.2
23-1	91.4		39-5	89.4		51-4	98.0
23-2	94.1		39-6	94.8		51-5	93.0
25-1	98.4		39-7	91.3		51-6	93.3
25-2	97.7		39-9	92.4		51-7	89.3
25-3	96.4		41-1	98.9		51-8	91.7
25-4	98.8		41-2	98.0		51-9	93.5
25-9	96.7		41-3	97.8		53-1	96.9
27-1	98.3		41-4	99.5		53-2	94.8
27-2	92.9		41-5	92.0		53-3	90.5
27-3	98.5		41-9	97.4		53-4	93.9
27-4	98.1		43-1	97.6		53-5	87.2
29-1	97.4		43-2	94.7		53-6	89.3
29-2	98.2		43-3	98.4		53-7	90.1
29-9	94.5		43-4	98.5			
31-1	94.6		43-5	95.8			

Source: OECD calculations based on Burning Glass data.

Annex B. The O*NET+ taxonomy

Table B.1. Construction of the O*NET+ taxonomy

Broad category	Category label	Source
Attitudes	Adaptability/resilience	ESCO
	Motivation/commitment	ESCO
	Self-management/rigour	ESCO
	Work ethics	ESCO
Arts and Humanities	Fine Arts	O*NET
	History and Archaeology	O*NET
	Philosophy and Theology	O*NET
Business Processes	Clerical	O*NET
	Sales and Marketing	O*NET
	Customer and Personal Service	O*NET merged
Production and Technology	Telecommunications	O*NET
	Building and Construction	O*NET
	Engineering, Mechanics and Technology	O*NET
	Design	O*NET
	Food Production	O*NET
	Production and Processing	O*NET
	Transportation	O*NET
	Equipment Selection	O*NET
	Quality Control Analysis	O*NET
	Installation and Maintenance	O*NET merged
Medicine	Medicine and Dentistry	O*NET
	Psychology, Therapy, Counselling	O*NET merged
Law and Public Safety	Law and Government	O*NET
	Public Safety and Security	O*NET
Science	Biology	O*NET
	Chemistry	O*NET
	Geography	O*NET
	Physics	O*NET
	Sociology and Anthropology	O*NET
Industry Specific Knowledge	Industry Specific Knowledge	New category
Languages	Local language	O*NET
	Foreign language	O*NET
Physical Skills	Psychomotor Abilities	O*NET
	Auditory and Speech Abilities	O*NET
	Visual Abilities	O*NET
	Physical Abilities	O*NET merged
Cognitive Skills	Originality	O*NET
	Quantitative Abilities	O*NET
	Reasoning and Problem-solving	O*NET merged
	Learning	O*NET merged

Communication	Active Listening	O*NET
	Reading Comprehension	O*NET
	Speaking	O*NET
	Writing	O*NET
	Communications and Media	O*NET
Digital	Office Tools and Collaboration Software	New category, based on ESCO
	Digital Content Creation	ESCO
	Digital Data Processing	ESCO
	ICT Safety, Networks and Servers	New category, based on ESCO
	Computer Programming	O*NET
	Web Development and Cloud Technologies	New category
Resource Management	Time Management	O*NET
	Management of Material Resources	O*NET
	Management of Financial Resources	O*NET merged
	Management of Personnel Resources	O*NET merged
	Administration and Management	O*NET
Social Skills	Coordination	O*NET
	Persuasion and Negotiation	O*NET
	Social Perceptiveness	O*NET
	Judgment and Decision Making	O*NET merged
Training and Education	Training and Education	O*NET

Source: OECD elaborations on O*NET and ESCO hierarchies.
Note: The mention "O*NET merged" in the last column indicates when the category is the result of merging two or more O*NET original categories.

Annex C. More details on the BERT model

BERT Base, BERT Large, RoBERTa and other language models

110. A number of other versions of BERT and other language models are available, and this Annex justifies the choice of using BERT Base rather than other versions or other models.

111. The original developers of BERT released two versions: BERT Base and BERT Large. The latter is a larger and heavier version of the first. It achieves slightly better results on the standard benchmarks than the smaller version. However, its size makes it significantly harder to optimise: it is longer to train and needs more memory (which usually means making sacrifices on other parameters). This also translates into lower transferability and reproducibility.

RoBERTa was proposed by Facebook AI in 2019 (Liu et al., 2019[35]). It relies on the exact same architecture as BERT but uses different hyper-parameters and training objective. The researchers affirmed that BERT was undertrained, i.e. did not reach the optimal performance its architecture allowed. On the classification of Burning Glass skills, RoBERTa did train faster (shorter time to reach a given threshold of accuracy) but achieved similar or slightly poorer final results.

112. A number of subsequent language models came after BERT since 2018: XLNet, DistilBERT, ALBERT, GPT2, and many more, each of them claiming either better performances or better training efficiency. While BERT is not the best performing model on the standard NLP benchmarks anymore, it remains the best known and documented in the field, and it makes for an uncomplicated and parsimonious (in computing power) approach.

Tokenisation

113. BERT does not actually build as many embeddings as words in the input sentence. If it were the case, BERT would need to store an initial embedding for each possible word it might encounter (including conjugations, plurals or spelling mistakes). However, this would mean storing several millions of vectors for several millions of words, which would be impractical and slow to use. Furthermore, for a particularly rare word, the quality of the embedding could be poor.

114. Instead, BERT resorts to tokenisation, namely breaking a word into sub-pieces called tokens, a traditional approach in NLP. BERT uses WordPiece tokenising. The methodology relies on splitting a word into its prefixes, root and suffixes. For example, “rare” is conserved as [“rare”] but “rarer” is split as [“rare”, “##r”] where “##” is used to distinguish between the suffix and the sole letter “r”. When processing a sentence including “rarer”, the Transformer architecture is in charge of spotting both the root and the suffix and update the corresponding embeddings accordingly in the subsequent layers. With this tokenising method, BERT reduces the number of unique tokens to be stored to around 30 000. For more details, see Schuster and Nakajima (2012[55]).

115. In practice, for each skill, there are as many embeddings as tokens in the definition, a number always greater than the number of words themselves. The maximum number of tokens that BERT can process at once is 512, but because of computing power issues, the taxonomy algorithm uses a maximum number of tokens equal to 256.

Annex D. Additional evaluation metrics

116. Accuracy is usually the primary metric of interest for a final user. However, in the presence of unbalanced categories, it can be misleading: predicting systematically the predominant category can yield high accuracy, but low predictive value. To counterbalance that, several other metrics exist: precision, recall, F1 score, Matthew's Correlation Coefficient (MCC), and many others. While those are quite standard and well defined for binary classification, their extension to multiclass classification is not natural. The precision, recall and F1 score are generalised by transforming the problem into a binary classification (correct/incorrect prediction) for each class and averaging the results over all classes. However, the averaging can or cannot take into account class imbalance.[28]

- The macro averaging calculates metrics for each class, and finds their unweighted mean. This does not consider label imbalance.
- The micro averaging calculates metrics globally by counting the total true positives, false negatives and false positives.
- The weighted averaging calculates metrics for each label, and finds their average weighted by support (the number of true instances for each label). This alters 'macro' to account for label imbalance (it can result in an F1 score that is not between precision and recall).

117. The "micro" definition of precision, recall and F1 score is equivalent to accuracy. Therefore, "weighted" scores are reported. The MCC is more easily generalised (Gorodkin, 2004[56]). A MCC of 1 indicates perfect correlation, whereas a 0 means "no better than random".

118. The table below summarises those metrics:

Table D.1. Additional evaluation metrics

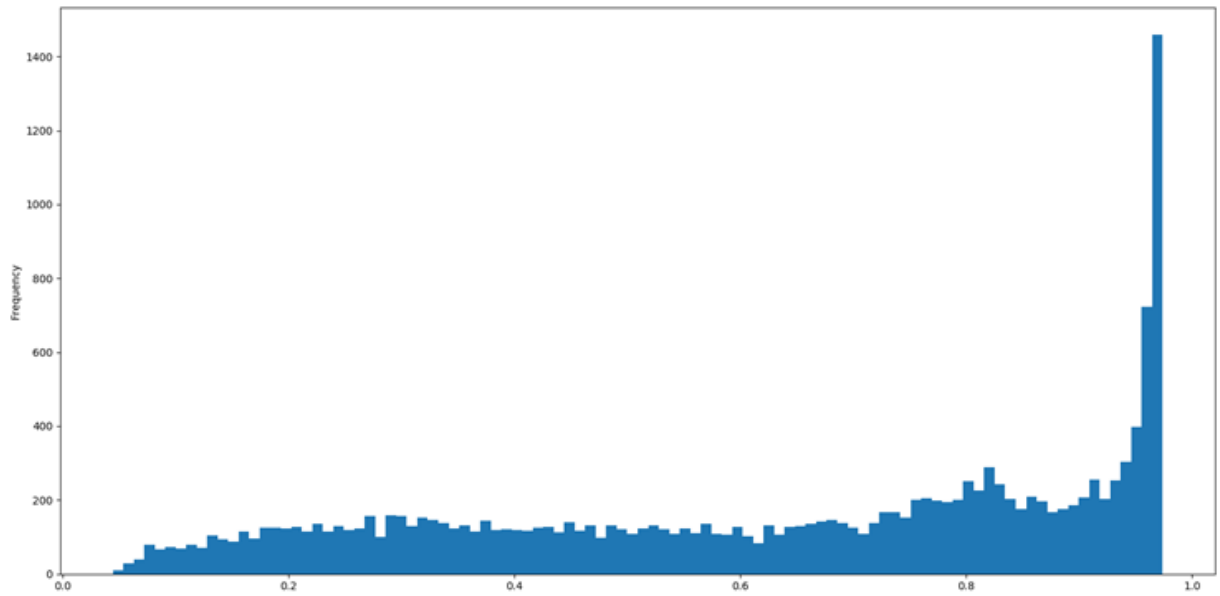
	Accuracy	Precision	Recall	F1 score	MCC
BERT (O*NET+ with 61 categories)	74 %	0.768	0.740	0.737	0.725
Random	2 %	0.058	0.017	0.022	0
Zero Rule	14 %	0.020	0.140	0.034	0
BERT (O*NET+ with 16 categories)	82 %	0.843	0.825	0.828	0.796

Source: OECD calculations based on Burning Glass data.

[28] From scikit-learn documentation

Annex E. *q* distribution

Figure E.1. Distribution of the q value over all classified Burning Glass skills

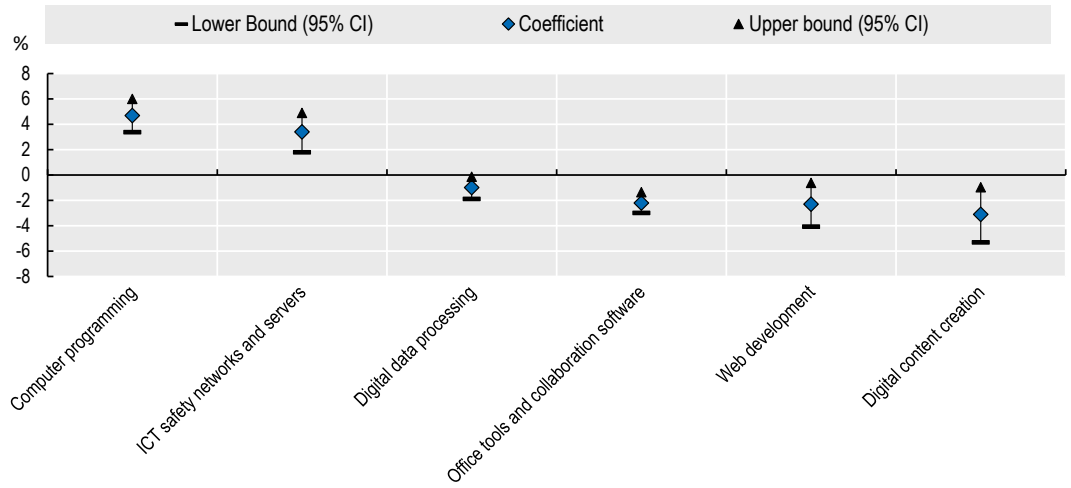


Source: OECD calculations based on Burning Glass data.

Annex F. Further Applications

Figure F.1. Hourly wage elasticity by skill requirement: Digital skills subcategories

Percentage change in hourly wages if the posting requires at least one skill in the category, everything else held constant.



Note: The graph plots the same specification as Fig. 12 (including all the same regressors) but the dummy for "Digital" skills, which was substituted with the six dummies for the subcategories of "Digital", as reported in the present figure. Canadian job postings data for 2018.
Source: OECD estimations based on Burning Glass data for CAN (2018).