Statistical Journal of the IAOS 40 (2024) 591–602
DOI 10.3233/SJI-230110
IOS Press

Enhancing taxonomy-based extraction: Leveraging information from online community platforms for digital skills demand identification in job ads

Joanna Napierała*
The European Centre for the Development of Vocational Training (CEDEFOP), Thermi, Greece

Abstract. The rapid technological changes have revolutionised how we function, including how we search for work and what skills we need to be equipped with to perform the tasks at the workplace. As employers more often recruit using online job advertisements, their content becomes a natural source of information for analytical purposes on the skills demanded in the labour market, especially for analysing emerging skills like digital. There are still some challenges with the extraction of information from online content. However, the extraction improvements go hand in hand with new technological developments like natural language processing techniques. This article presents the experimental method of updating the classification of digital skills to keep it up to date for information extraction applied to online job advertisements. The evaluation proved this method successfully identified terms related to programming skills but failed to identify terms associated with artificial intelligence sufficiently. The latter is related to the fact that the AI field is among the fastest developing areas of technology advancement, and new terms (e.g. Chatgpt) always appear.

Keywords: Digital skills, skills extraction, online job advertisements, big data, skills intelligence

1. Introduction

Apart from the demographic changes, namely the ageing of the population, the situation on the labour markets in the countries of the European Union is nowadays shaped by two significant trends: the greening of the economy (e.g. companies implementing circular economy approach) and the digital transition linked to the processes of digitalisation, automation and robotisation. As the so-called twin transition is happening at an unprecedented speed, the traditional sources of labour market information (e.g. labour market surveys) may need to provide timely information to identify these thorough changes in the structure of demand for workers' skills. The twin transition is also very disruptive, meaning that it is not only changing how the workers perform their tasks and what skills they need to accomplish these but also is causing the whole workplace to become obsolete and eventually disappear, increasing the poll of workers who require reskilling to take up new employment. Therefore, the analysis based on non-traditional data, such as the content of online job advertisements (OJAs), is gaining momentum (see [1]). The advantage of using this kind of data, characterised by high granularity and near real-time availability, is that it may provide all stakeholders, including labour market and education system policymakers and training providers, with valuable support by bringing a better understanding of the directions of these expected changes.

Yet, before this kind of support based on online job advertisements will be achieved, the challenges related to information extraction must be addressed. It has been

*Corresponding author: E-mail: joanna.napierala@gmail.com.

shown [2,3] that information retrieval solutions based on structured and fully semantic ontological approaches or classifications work better, allowing for meaningful information extraction. However, such high-quality training datasets are rarely available. Moreover, the frequency of the revisions of existing classifications will always interfere with the quality of information extractions. In particular, the lack of regular updates in classifications becomes problematic for analysis of the skills emerging in the labour market with the implementation of innovations or new technologies (e.g. new software, patents). The focus of this article is to present the evaluation results of a method proposed to address the challenge of keeping the classification of digital skills up to date. This method was developed under the Cedefop project: "Towards the European Web Intelligence Hub – European system for collection and analysis of online job advertisement data (WIH-OJA)".

This article is structured as follows. In the opening section, we discuss the significance of online job advertisements as a valuable source of information for labour market analysis. We emphasise the relevance of this source of information in gaining insights into the changes in demand in the labour market, particularly concerning the skills required. The second section delves into the intricate challenges of extracting skills information from online job advertisements. This section discusses the developments in identifying and categorising skills as they appear in the content of job advertisements. The third section is dedicated to the proposed solution to ensure that digital skills classification used for information extraction from online job advertisements remains up-to-date and relevant. Here, we also provide a detailed examination of the specific challenges encountered when extracting digital skills from online job advertisements with this method. In the final section, we estimate the demand for digital skills in the European Union to evaluate the improvements brought by the proposed method for information extraction.

2. Online job advertisements as a source of information

A few decades ago, when job advertisements became a valuable source of data for researchers, their use for analytical purposes was primarily motivated by a growing interest in understanding the evolving nature of skills demanded in the workplace [4]. Previously, job advertisements were spread out via word of mouth or communicated via printed media (e.g. specialised journals). Still, more recently, the Internet of Things and digitalisation have changed how individuals behave, including how employers recruit and people search for work. Consequently, the recruitment has moved to specialised sites and social media platforms. This and rapidly advancing natural language programming techniques have considerably changed how researchers can access and analyse information from job advertisements, making it much easier to analyse demand in the labour market. However, what did not change is that this information reflects the skills and qualifications employers seek in their future employees, allowing researchers to gain insights into the types of jobs or skills newly created in the labour market.

The timeliness or level of detail is an unquestionable strength of OJA as a source of information about demanded skills (e.g. [2]). Yet, the skills extraction approach based on OJAs also has some weaknesses. The major flaw comes from not all vacancies being advertised online. The analysis of the coverage shows that despite observing an upward trend, as after Covid19 pandemic, far more employers than before are using online recruitment to reach out to potential candidates; still, the low-skilled professional occupations tend to be under-represented in the data. Also, larger enterprises have more preference to use this channel for recruitment compared to small ones (62% vs. 29%) and the recruitment occurring in urban areas will be better reflected in this source of information than in rural regions (see [5]). Apart from coverage bias, duplication of the same advertisement is another challenge to overcome. The ease of posting online can encourage employers to use different web portals to increase the pool of potential candidates. Alternatively, some web portals, so-called aggregators, apart from collecting their job advertisements, also include links to information from other websites. To avoid overestimating the number of vacancies, the careful choice of sources and the deduplication process is necessary for every data pre-processing system when analysing job postings [6]. Some challenges in this type of analysis are also related to advertisement content. For example, employers may deliberately omit some skills when drafting OJA content, especially when their possession is implicitly expected from candidates (e.g., computer skills from IT professionals). This could lead to an underestimation of the number of jobs that require such skills based on OJA analysis. For example, the estimation of basic digital skills derived from mentions of basic computer literacy (2.9% of all advertisements) or the use of office software (11.3%) was perceived as lower than expected in the labour market [7].

Apart from the challenges, online job advertisement data, compared to more traditional sources of information like surveys or administrative data, offers access to nearly real-time information. In the rapidly evolving labour market, the pace of changes requires the labour market or education policymakers to build their knowledge based on the results of analysis of data sources offering the timeliest information. Therefore, despite some flaws, the content of online job advertisements as a source of information has recently gained a lot of attention in analysing the impact of trends, such as robotisation, automation or green transition, on the changing labour market and skills (see [1]).

3. Extracting skills from online job advertisements

Job advertisements typically contain standard information such as job title, location, employer name, job description, qualifications, and other skill requirements. While it's relatively easy for humans to extract relevant information from the free-text content, it remains challenging for machines [8]. The unstructured data requires preliminary cleaning before existing machine learning methods can utilise it. This is especially true since many web portals do not impose specific structures, leaving decisions on content length and the information included to the individuals posting the advertisements. It is, therefore, self-reported information employers use to inform potential candidates about the skills required for the tasks performed at the future job.

In the past, the "bottom-up" approach was more common as the extraction of information method, with text being manually or semi-automatically identified as words and phrases related to skills and competencies and later treated with clustering or topic modelling techniques for grouping similar concepts. For example, Loth et al. [9] manually tagged text extracts from the sample of 200 advertisements to distinguish 13 information types (e.g., occupation, company, sector, contract, competence, personality, education). Gao and Eldin [10] applied the Bayes classifier first to assign each sentence extracted from an online job advertisement to one of two groups: either containing job qualification information or not. Later, the unsupervised approach for topic modelling, namely Latent Dirichlet Allocation (LDA), was used to identify the groups of skills (Ibidem). In this setting, the sentences, including requirements about skills, were treated as the documents and the required skills as topics to be discovered.

However, with technological advancements in automatic text classification, the "top-down" approach became more common as it proved to perform more efficiently than the supervised "bottom-up" approach, which needs costly and time-consuming expert annotation [2,3]. In particular, regular expressions, now part of the standard library of many programming languages, including Java and Python, were helpful for pattern matching, allowing one to identify or extract particular pieces of text from a string or document.

The 'top-down' approach can be described as the task of using machine learning techniques in assigning natural language texts to predefined categories (e.g. European Skills, Competences, Qualifications and Occupations[1] (ESCO) or the Occupational Information Network[2] (O*NET) in the case of skills detection or International Standard Classification of Occupations[3] (ISCO) when classifying the job titles). For example, word embedding algorithms are first applied to associate words, word forms or phrases with ESCO skills terms based on vectors whose similarity is compared using Levenshtein distance, Jaccard similarity, and the Sørensen-Dice indexes. Once a pair of terms had higher similarity than 70% and was accepted by domain experts in their manual evaluation, it was finally admitted to be used in skills extraction from online job advertisements (see [11]). In the literature, other similarity measures (e.g., Cosine, Motyka, Ruzicka) were applied to evaluate the matching between extracted and taxonomy-based terms. In the top-down approaches, the quality of extracted information about skills tends to be as good as the underlying taxonomies used for this purpose [12]. Therefore, one of the limitations of using taxonomies is that they may not be comprehensive and can quickly fall out of date as new skills are constantly emerging. Therefore, some researchers decided to build their taxonomies, e.g. [13] created a taxonomy of soft skills based on DBpedia. In the study of job advertisements in Austria, Plaimauer [14] showed that more than half of the ESCO skills terms have never appeared in the content of online advertisements, also pointing to the length of term as being reversely associated with the frequency of their appearance in the descriptions. Also, the grammatical cases in some languages seem challenging for natural language processing tools, often leading to misinterpretation of recognised skills [15].

[1] https://esco.ec.europa.eu/en.
[2] https://www.onetonline.org/.
[3] https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/.

4. Analysis of digital skills using online job advertisements

Digital technologies such as artificial intelligence, the internet of things, cloud computing, and recently also large language models such as Chatgpt are the main forces behind reshaping the world of work nowadays. The increased interest in understanding the demand for digital skills is related to their growing importance for employers. For example, the use of Austrian and German longitudinal data confirmed that, among others, the increase in the use of social media in the workplace of public relations specialists translated into higher demand for digital skills among this occupational group [16]. As individuals' skills are essential in making any change happen, it is important to understand trends in labour market demands professions leading the digital transition, including software developers, ICT technicians, and programmers (e.g. [17]). For example, the analysis of advertisements in Germany showed that knowledge of programming languages such as Java, C and Python was in most demand for computer scientist positions [18].

Analysing job offers also allows one to understand emerging occupations, which did not exist when the taxonomies, such as the ISCO, were updated in 2008 [19]. The comparison of advertisements and skills requested for ICT and statisticians' roles confirmed that making distinct classifications of these two occupations still holds, e.g., computing skills are distinctive for software developers [20]. Analysing skills contained in online job advertisements may have the final purpose of informing education providers more efficiently, such as the work of Gurcan and Cagiltay [21], which contributed to gaining more insights about competencies that big data software engineers are expected to have.

Another strand of literature uses online job advertisements in analysis to understand changes in demand for skills for occupations that transform due to technological change. Acemoglu et al. [22] introduced three ways of estimating exposure to artificial intelligence at occupational levels. They used longitudinal online job advertisement data for the US to show how demand for workers with AI skills has increased over time. The recent work of Anton et al. [23], in which OJAs from 60 countries were analysed, contributed to a better understanding of the competencies required for leveraging AI. In June 2023, the number of employers mentioning Large Language Models (e.g. ChatGPT) among required skills was negligible in the labour market in Europe [24]. Yet, a similar analysis of the information from China's largest online recruitment platform brought insights into the employers' requirements for this skill, predicting that 45% of all future occupations might require it [25].

Online job advertisements were also used in discussions of the COVID-19 pandemic's impact on the labour market [26], which triggered a higher prevalence of working from home associated with a higher demand for digital skills. Namely, the basic digital skills of workers expected to work using digital solutions and the high-digital skills of IT workers securing the connections and being responsible for running the IT system of firms. Indeed, even after the end of the pandemic, basic digital skills, e.g. computer use, were twice as likely to appear in the content of advertisements whenever the possibility of working from home was explicitly mentioned [27].

5. Description of the method to update digital skills classification

The increasing interest in the content of online job advertisements as the source of information made the European Union Agency – Cedefop, join forces with the statistical office of the European Union – Eurostat, in developing the European system for collection and analysis of such data under the umbrella of Web Intelligence Hub project. This so-called WIH-OJA database, which is used in the analysis presented in this article, is based on the content extracted from online job advertisements (OJAs) published on various websites selected by national labour market experts in the way to assure good coverage of diverse labour markets segments in EU (e.g. including websites which do not overlap in the offer of same services). This joint project uses the European multilingual classification – ESCO – to classify information extracted from online job advertisements.

A good taxonomy of skills, not just digital ones, is crucial for many education system stakeholders, including learners, employers, and policy-makers. A clear taxonomy can guide curriculum development for educational institutions and training providers, ensuring that all essential skills are covered and a logical progression in the learning journey. Individuals can use the taxonomy to identify gaps in their skillset, helping them make informed decisions about further training needs to progress in their careers. For companies, a good taxonomy is a tool that could help in recruitment by specifying exactly what skills are required for a given role. As

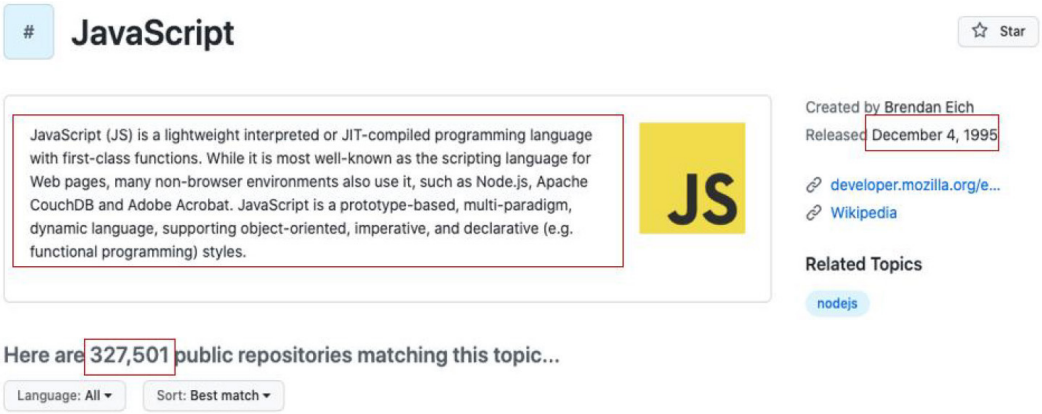

Fig. 1. Illustration of the information about tag "Javascript" obtained from GitHub page.

digital technologies change rapidly, for the taxonomy of digital skills to remain relevant, it must be updated regularly to reflect these changes.

The most recent version of ESCO, released at the beginning of 2022 (version 1.1.1), includes 13890 skill terms and knowledge concepts, out of which 1,201 are labelled as digital. However, when the experimental project described in this article was initialised, the initial list of digital skills available from ESCO counted only 21 skill terms linked to the Digital Competence Framework for Citizens[4] (DigComp). This experimental project aims to test if adding an alternative list of terms parallel to the existing ESCO terms would improve the extraction of information about digital skill terms. The potential of using tags obtained from the Stack Overflow.com website was explored to build a list of the latest IT technologies, which were used to compare with the content of OJAs.

Although a list of digital skills was updated recently, the simple comparison of Stack Overflow survey results shows that half of the 45 most commonly used programming languages (e.g. Bash, Shell, Clojure, Crystal, Dart, Delphi, Elixir, Fortran, Julia, Kotlin, Lua, SAS) in 2021 are missing in the ESCO classification. This task aimed to explore the potential of using information from the Stack Overflow platform to provide regular updates for the classification of digital skills.

The Stack Overflow is a community-based platform[5] anyone can access and use to pose a question, find an answer, or contribute a solution to technical challenges. The popularity of this public platform is confirmed by a high number of users (100 million each month on 6/10/2022), mainly software developers, who collectively contribute to building knowledge and using this platform for information exchange. The richness of the information in this database has attracted the attention of researchers in the past who used this data to analyse various aspects [28,29]. Yet, to the author's knowledge, this source of information was not previously used to extract a list of technologies to help update the existing classification of digital skills.

The main benefit of using Stack Overflow data for this purpose is that each question or answer is categorised by receiving a tag that describes the topic of the question and allows for linking and grouping of similar discussion topics. For example, any question, as shown in Fig. 1, should receive a unique tag, i.e. #JavaScript, because the query relates to the programming language JavaScript.

The tags were introduced to increase the platforms' functionality and improve the easiness of their navigation. At the time of project realisation, the most popular tag for #javascript had more than two million recorded questions. The platform also allows the creation of a list of synonyms or alternative labels for the same tag. For this project, it was helpful to use the list of Stack Overflow tags together with the information about the number of times platform users have used them as a proxy of their popularity to create a list of potential IT technologies crucial for employability. The first extraction of information from the Stack Overflow platform gave us back more than 65ths of tags that were used at least once by any of the users of this platform since its opening in 2008. Yet, not all of them were relevant in the context of information extraction about digital skills. To narrow down this number of tags, by excluding technologies of

[4]https://joint-research-centre.ec.europa.eu/digcomp_en#:~:text=The%20Digital%20Competence%20Framework%20for,and%20for%20participation%20in%20society.

[5]https://stackoverflow.com.

tags	description_final	synonyms_string	num_question_stackoverflow	num_question_git	release_date
javascript	JavaScript (JS) i...	['js', 'ecmascrip...	2387099	322832	12/04/1995
python	Python is a dynam...	['pythonic', 'pyt...	1968596	287713	02/20/1991
java	Java was original...	['java-se', '.jav...	1851865	169544	05/23/1995
csharp	"C# (pronounced "...	[]	1543060	48325	01/01/2002
php	PHP is a popular ...	['php-oop', 'php-...	1439082	91702	06/08/1995
android	Android was desig...	['android-mobile'...	1379107	98332	09/23/2008
html	HTML, or Hypertex...	['html-tag', 'htm...	1135474	147198	06/01/1993
jquery	jQuery is a light...	['jquery-core', '...	1030600	31917	01/01/2006
cpp	C++ is a popular ...	[]	767980	47078	10/01/1985
cxx	C++ is a general-...	[]	767979	434	01/01/1900
css	Cascading Style S...	['cascading-style...	762382	157875	12/17/1996
ios	iOS is the operat...	['iphone-os', 'ap...	671397	37268	06/29/2007
mysql	MySQL is an open ...	['my-sql', 'mysql...	649023	42716	05/23/1995
sql	"SQL stands for s...	['sql-query', 'sq...	633663	25535	01/01/1986
r	R is a free progr...	['rstats', 'r-lan...	452552	26768	08/01/1993
node.js	Node.js is a tool...	['nodejs', 'io.js']	432609	166176	05/27/2009
arrays	An array is an or...	['array', 'array-...	396277	1856	01/01/1900
react	React (also known...	['react', 'react-...	394832	218706	03/01/2013
c#-nameof	C is a programmin...	[]	381082	45245	01/01/1972
asp.net	ASP.NET is an ope...	['asp-net', 'aspx...	368725	733	01/01/2002

Fig. 2. The excerpt from the final list of IT technologies used for information extraction.

minor relevance to the current labour market situation, the information from Stack Overflow was further complemented with similar details about the frequency of using this tag from another platform used by millions of software developers, namely GitHub.com. 80% of the most frequently asked questions on GitHub included 580 tags, while in Stack Overflow, 80% was equalled to 1300 tags. Because the tags from GitHub matched with those provided by Stack Exchange. Eventually, the 1300 tags were used for information extraction from OJAs.

6. Challenges in information extraction about digital skills

Application of the 1300 tags list obtained from the Stack Overflow platform to annotate the content of OJAs revealed some challenges in using it for data extraction. The first challenge is related to the presence of terms. Recruiters may not mention some essential skills. For example, as “Drupal” is a free and open-source web content management system written in “PHP”, someone can assume that the knowledge of “PHP” suffices to be mentioned in OJAs to recruit a person who will be capable of using Drupal. Nevertheless, looking at terms that were missing in OJAs but present in skills classification could help identify the IT technologies that have become obsolete. For example, the knowledge of programming languages like Pascal, designed in 1969, or Prolog, developed in 1972, which became replaced by other programming languages in the meantime, may have disappeared from OJAs for such reason. Conversely, looking at the list of programming languages requested by employers and not included yet in ESCO classification, we could identify new relevant skill terms. For example, Golang (sometimes termed Go Programming Language) would be one of them.

The second challenge is related to the fact that the names of IT technologies are usually one-word terms with double meanings. For that reason, when used for information extraction, they might be confused with other non-relevant terms in the context of digital skills. Therefore, the inclusion of new terms to the existing classification needs to be proceeded with the evaluation by human experts and the development of training data sets. For example, the high detection of demand in OJAs written in a specific language compared to overall demand might indicate that the skill term has a double meaning. Specifically, the term “lua” means “to contact” or “to get” in Romanian and was detected in OJAs written in Romanian more frequently compared to other languages. The same regarded term, “Sprache”, which is a lightweight library for constructing parsers directly in C# code, but in German, this term also means “language”. It could be easily confused by algorithms with requirements about the ability to speak foreign languages.

Furthermore, cross-checking the data at the occupational level may reveal other problems related to a double meaning of some terms. For example, the term “Rails” used for tagging questions related to a web application framework written in Ruby may lead to incorrectly identifying IT skills for the occupation of plasterers. In the OJAs recruiting for roles as plasterers, we may find the word “rails”, which does not refer to digital skills (e.g. “we search for a person who knows how to install rails, smooth walls, etc.”). Similarly, tag “Lager”

is not only a type of beer but also a logging framework for Erlang, in which case the digital skill could be mistakenly annotated for the occupations related to brewery as workplace. Another example is the operating system called "Pick". Analogously to the example with "lua", the term "pick" can also be found in other contexts than as a reference to the knowledge of the system, e.g. seasonal occupations in agriculture where picking of fruits/vegetables will be required. The term "Jacket" could either point to the requirement of knowledge of numerical computing platforms enabling GPU acceleration of MATLAB-based codes or could come from OJA recruitment of someone possessing tailoring skills. The term "Rivets" is a C++ class library which provides the infrastructure and calculational tools for particle-level analyses for high-energy collider experiments. But if it is used in a non-digital context, it could mean the physical skill of an assembly cell operator or welding operator. Finally, the tag "Intern", a JavaScript testing framework, could be confused with vacancies targeting recruitment of interns.

Another challenge concerns some tags that could be confused with existing company names. For example, Whirlpool (sometimes styled WHIRLPOOL) is a cryptographic hash function that transforms any arbitrary block of data into a fixed-size string of bytes that is unique (to the best of human knowledge) to the given data. But this term is also the name of a multinational manufacturer and marketer of home appliances. Similarly, using SPA could detect demand for well-being and spa resort workers instead of identifying skills related to web application – single-page application (SPA). The application of a technique called Named Entity Recognition (NER), a subfield of Natural Language Processing (NLP) that involves the identification and classification of named entities (like names organisations), should help in the correct distinction between the digital skill term and the name of a company.

On top of terms causing problems in specific languages or ambiguity depending on the occupation, we have also encountered one-digit words used to describe IT technologies, which will be impossible to extract as these are by default cleaned by tools that remove stop words, sparse terms, and similar particular words before the content of OJAs is being processed. One example is the programming language "R", a free tool often used by researchers, statisticians, or data scientists.

A related challenge to one-digit words occurs with using some tags, which are abbreviations. For example, the acronym DFA could mean either Deterministic Finite Automation, a simple model of computation, or the Department of Foreign Affairs or Financial Analysis Department, which are unrelated to digital skills.

The term's ambiguity could also be related to the context in which it was used. For example, knowledge of PDF editor, if extracted based on the presence of the term "PDF", could lead to mistakes when recruiters, instead of looking for candidates proficient in the use of PDF, encourage them to "upload" their CV in PDF format to apply for the position. A similar mistake could be made with all skills related to social media use; therefore, the machine learning model needs to be trained to distinguish between requests for using Instagram and developing content on Instagram from the mention of Instagram itself or other social media tools.

Some terms could be used for information extraction but only if they are accompanied by another word, e.g., "warehouse" itself most likely will lead to the identification of jobs in a warehouse but, when searched jointly as "data warehouse" instead, could help in identification of digital skills. Similarly, the word "tracking" could only make sense if accompanied, e.g. "video tracking" or "online tracking"; the word "virtual" together with "reality" would make sense as otherwise could mistake roles in a virtual working environment that refers to the possibility of remote work.

A simple matrix showing frequencies of found terms by language and occupations will allow identification of ones appearing more often. This, in turn, will indicate potential problems with the ambiguity of the meaning of terms. As shown in Fig. 3, the presence of the term 'LUA' in OJAs written in Romanian for the data entry clerk was relatively high, but as it was not present neither in OJAs written in English or German language, we may assume that in that case, this extraction is wrongly indicating the demand for digital skill for this occupation. Similarly, the term 'Sprache' was frequently observed in job advertisements in exemplary occupations. Yet, the demand for Sprache was not observed in either of these occupations in advertisements written in English or Romanian. Therefore, we may also assume that in these cases, the employers searching for either nannies or delivery drivers referred to knowledge of languages rather than requesting digital skills. Such a thorough cross-check of terms applied across all languages and occupation levels will allow for building up a list of cleaning rules for algorithms used for data extraction, which would exclude these terms as invalid ones for these occupations. By setting up a cleaning rule, it is understood that for any problematic combination of language, occupation and an ambiguous term, the extracted term will not be classified as a valid digital skill term and will be left out.

Name of occupation	presence of term per thousand of OJAs					
	LUA			SPRACHE		
	de	en	ro	de	en	ro
data entry clerk	0	0	100	88	0	0
medical device engineer	0	0	8	26	0	0
car and van delivery driver	0	0	8	38	0	0
nanny	0	0	4	63	0	0
administrative assistant	0	0	3	54	0	0
warehouse operators for clothing	0	0	3	56	1	0
set builder	0	0	2	52	0	0
sprinkler fitter	0	0	2	22	0	0
publishing rights manager	0	0	1	53	0	0
bricklayer	0	0	1	28	0	0

Fig. 3. The terms LUA and Sprache per thousand OJAs by language and occupation. Source: WIH-OJA.

For terms which could be classified differently according to the context, for example, a demand for Sprache in the case of IT occupations, which could either refer to a demand for knowledge of language or digital skills, dedicated training datasets need to be developed. A comprehensive set of advertisements that includes the term in various contexts needs to be gathered, covering as many potential meanings and uses of the term as possible. Human experts must assess each instance of the term within its context and annotate them manually, distinguishing between cases regarding the demand for digital skills or another case. The dataset with annotated terms must include a balanced representation of the different meanings to ensure and prevent model bias towards the more commonly occurring context. The annotated terms can train a machine learning model to distinguish between ambiguous situations and correctly classify the term. After the model's performance evaluation, it can be deployed in a real-world environment. Having both cleaning rules and training datasets in place will eventually allow for making an informed selection of terms valid for extracting digital skill terms from the content of OJAs.

7. Digital skills demanded in EU27

Digital skills generally encompass a broad range of capabilities related to effectively using digital devices, software, and platforms and creating, using, and understanding digital content. Programming skills, a subset of digital skills, refer to the ability to write and understand computer code to create software programs, scripts, or other sets of instructions for computers to follow, are considered. The worldwide survey results from 2022 show that respondents, who were mainly professional developers (75.39%), used for work 45 different programming languages.[6] The comparison of the mentioned names with the existing skills terms in ESCO taxonomy version 1.0.1 revealed that out of the 45 programming languages, 24 were already included in the taxonomy. Additionally, Fortran, Lua, OCaml, PowerShell, and Bash also existed in ESCO taxonomy as alternative labels of generic skill terms defined as computer programming. From the missing 15 skill terms, 12 were present in Stack Overflow as "tags". The programming languages not present in ESCO and Stack Overflow were C, F#, and Go.

The Stack Overflow surveys aim to explore developers' tools and technologies, identifying the most popular programming, scripting, and markup languages developers use. In 2022, the most popular ones were almost the same as in the previous year, namely JavaScript and HTML/CSS, followed by Python, which has overtaken SQL and Typescript. Yet, students reported using Python more often than SQL. Python was also the first language of choice for those who were not professional developers. Using the OJA data, in the first quarter of 2023, the two programming languages mentioned in one-quarter of advertisements targeting a group of technology professionals in the EU27 labour market were SQL and Javascript. Except for PHP, which ranked eleventh in the survey but was the third most popular demanded programming language by employers, the survey results align with the most in-demand programming skills in the EU27 labour market. Interestingly, in the observed period, the number of mentions with requirements of knowledge of Python has doubled from less than 5% to 10% (see Fig. 4). Also, the popularity of Typescript is growing. However, in the analysed period, it was still only observed in, on average, 1% of advertisements targeting technology professionals.

[6] https://survey.stackoverflow.co/2023/.

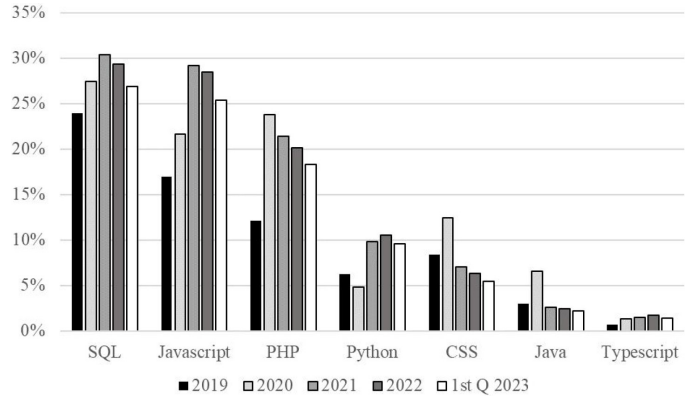

Fig. 4. Programming languages mentioned in online job advertisements targeting technology professionals in EU27. Source: WIH-OJA.

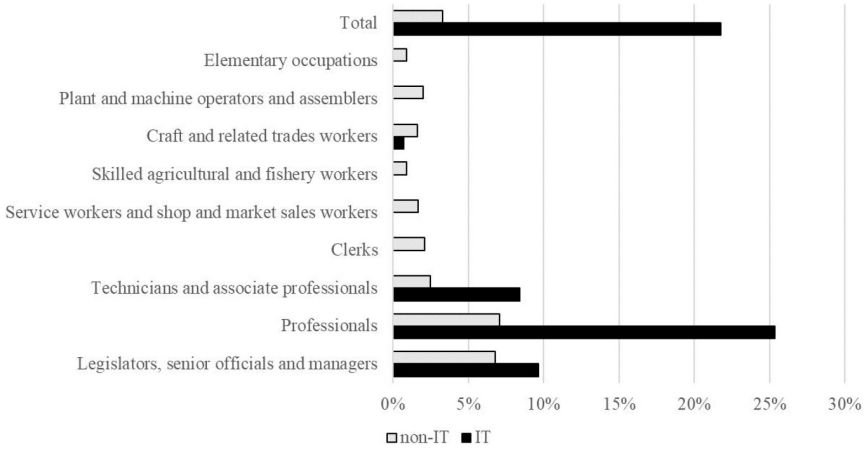

Fig. 5. The average difference in the share of advertisements requesting programming languages when extraction with ESCO classification was extended with 18 terms from Stack Overflow (in pp). Note: average calculated from 2021 and 2022 where 2022 data includes only observations until October. Source: WIH-OJA.

Determining the employer demand for 6th ranked programming languages Bash/Shell with ESCO classified OJAs was impossible as the "Bash" is hidden under the generic label of computer programming, and "Shell" was not included in the classification.

When utilising the 24 programming language terms from the ESCO classification as an indicator of overall demand for programming skills in 2021, 39% of job advertisements targeting IT professionals mentioned at least one of these terms (see Fig. 5). In contrast, only 1% of advertisements for roles outside of the IT sector did so. Adding 17 terms from Stack Overflow to complement the list of programming languages in identifying advertisements with demand for programming skills, we could see that these ratios increased to 52% for IT and 3% for non-IT occupations, respectively. When the data was broken down into occupational groups, the disparity in percentage points for IT occupations became more pronounced. Advertisements for legislators, senior officials, and managers saw a difference of nearly 25 percentage points, while IT workers in the craft and related trades category saw a three-percentage points difference. For the non-IT group, the highest difference was noted for the group of professionals, with a five-percentage point difference in estimating demand for programming skills.

Artificial intelligence (AI) skills are considered a subset of digital skills. They refer to the capabilities required to design, develop, implement, and maintain AI systems and solutions. Such skills can include expertise in machine learning algorithms, neural networks, natural language processing, robotics, computer vision, and other domains within the AI field. Detecting skills related to artificial intelligence in online job advertisements involves identifying keywords and phrases commonly associated with AI roles and responsibili-

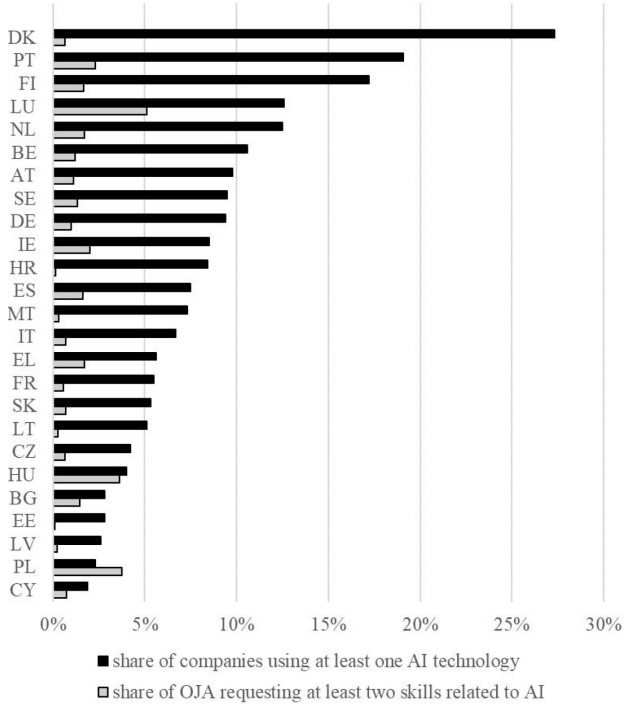

Fig. 6. The average share of advertisements requesting skills related to artificial intelligence. Source: WIH-OJA, ISOC_EB_AIN2.

ties. However, as AI is a rapidly evolving field, it is essential to continuously update this list based on the latest technologies and trends. To judge the ability of the Stack Overflow database to generate information about emerging AI skills terms, we have used a list of 210 terms proposed recently by the OECD as a benchmark [30]. In this list, the skill terms are grouped into seven clusters: Artificial Intelligence, Autonomous Driving, Machine Learning, Natural Language Processing, Neural Networks, Robotics, and Visual Image Recognition. The comparison of skill terms shows only a 20% overlap between Stack Overflow and OECD lists. Using the skills that belonged to the artificial intelligence cluster, where the overlap between Stack Overflow terms with the OECD list was the highest, with the same 11 out of 29 terms, we could estimate the demand for these skills in the EU. The demand for AI skills in OECD countries[7] was between 0.30% in 2019 and 0.40% in 2022 [30], whilst, using the terms extracted from the Stack Overflow database, the conclusion would be that the demand in the EU was much higher on average, 1.39% (see Fig. 6). Part of the high demand for AI skills is already related to the uptake of AI solutions on the market. However, higher demand for AI skills could also be observed in countries trying to catch up with the AI adoption by implementing the national AI strategies, e.g. Luxembourg.

The overall low overlap between Stack Overflow terms and the OECD list indicates that establishing the updates of the classification of digital skills solely based on information from Stack Overflow may not be sufficient to build an exhaustive list. This list could be extended by repeating the same exercise on platforms explicitly dedicated to Artificial Intelligence, e.g. AI Stack Exchange.[8] As a robustness check, the 1360 terms were downloaded from AI Stack Exchange[9] and tested if present in the content of a representative sample of online job advertisements targeting various IT professionals. The overlap between the AI Stack Exchange list of terms and Stack Overflow was only at the 11% level, indicating that the terms created by users of each platform differ. Yet, the overlap between AI Stack Exchange terms with the OECD list was at 28%, only eight percentage points higher than with Stack Overflow (see the Annex for the table showing the frequencies of overlapping terms). Apart from terms overlapping with

[7] Not all EU27 countries were analysed in OECD study.

[8] I would like to thank the anonymous reviewer for this valuable suggestion.

[9] State of the art on 2nd of January 2024.

Stack Overflow or OECD list, the other extracted terms had a very low presence in the OJAs (only 45 terms had frequencies above 1%). From the list of 45 terms which were present more frequently, few, e.g. "education", "cv", or "training", could refer to the steps in the application more than AI skills; few others, e.g. "automation", "features", "research" are too generic to indicate demand for AI-related skills definitely; few terms have already existed in ESCO taxonomy, e.g. "problem-solving"; finally only a few terms could be potentially considered as relevant. This indicates that the list built out of terms present at the AI Stack Exchange platform is also insufficient to increase the existing classification's capacity to extract AI skills.

8. Conclusions

The growing importance of digital skills in the workplace translates into the higher need for labour market analysts to understand what skills individuals need to be equipped with to find employment. The potential of using information obtained from OJA content is indisputable in this area, as information about occupations and skills required is very detailed. The systems of information extraction from unstructured text based on classification or taxonomy, such as the ESCO-driven approach shown in this article, tend to be more efficient in processing time and economic aspects, as they do not require costly post-validation of obtained information by experts. Yet, the main challenge of this approach is that the classifications or taxonomies, e.g., digital skills, become obsolete relatively fast. Exploring non-standard approaches to help keep the classifications up to date with emerging technologies that require new skills is necessary to make classification-based information extraction from OJAs relevant.

Based on extracting additional terms from the Stack Overflow platform, the method to provide regular updates to the existing classification of skills presented in this paper successfully identified approximately one thousand new digital skill terms. Yet, there are a few shortcomings of this approach. The first one is related to the fact that the language of information exchange on this platform is English. Although English is widely recognised as the lingua franca of information technology and many companies post job advertisements in this language, primarily when recruiting IT professionals [31], some terms might have been translated. Therefore, finding the equivalent translation of such terms to other languages is necessary to reduce the bias in extracting skills from advertisements not written in English. Nevertheless, the tested method was developed to support the enrichment of ESCO classification, which accepts proposals of skill terms in English, the working language of the group developing it. Once the terms are acknowledged to be included in the official version, the translation of terms is provided and validated by a network of national experts. The second shortcoming is insufficient coverage of the Stack Overflow platform to generate a good list of skills terms linked to artificial intelligence. As none of the three platforms tested, Stack Overflow, GitHub, and AI Stack Exchange, could provide a list of skills terms that would be exhaustive in this area, other methods need to be applied to complement it. It is worth mentioning that currently, the ESCO taxonomy is enriched on a bi-annual basis by collecting inputs from various sources, including the content of training offers, the content of curricula (based on Europass), the content of online job vacancies (based on EURES) and the stakeholders' inputs (e.g. HR or labour market experts), which jointly may allow for building the exhaustive list of new skill terms linked to AI. Other sources could also be considered, e.g. patent data. Eventually, an annual analysis of parts of OJA's content not classified with taxonomy could also bring back some new terms as input for ESCO enrichment. The third limitation, as the term's ambiguity depends on the context, is the overall challenge observed in information extraction from online job advertisements, which is not specific only to digital skills classifications. For example, the word Python may be related to the request for programming skills or data analysis skills using Python. Developing dedicated datasets with online job advertisement content annotated by experts is necessary to train machine learning models to interpret the terms accurately in various contexts. This is crucial to allow for successful information extraction.

Supplementary data

The supplementary files are available to download from http://dx.doi.org/10.3233/SJI-SJI230110.

References

[1] Napierala J, Kvetan V. Changing Job Skills in a Changing World. In: Bertoni E, Fontana M, Gabrielli L, Signorelli S, Vespe M, editors. Handbook of Computational Social Science for Policy. Cham: Springer. 2023. Available from: doi: 10.1007/978-3-031-16624-2_13.

[2] International Labour Organization. The feasibility of using big data in anticipating and matching skills needs. 2020. ISBN 978-92-2-032855-2.
[3] Sadro F, Klenk H. Using labour market data to support adults to plan for their future career: Experience from the careertech challenge. 2021.
[4] Harper R. The collection and analysis of job advertisements: A review of research methodology. Library and Information Research. 2012; 36(112): 29-54.
[5] Napierala J, Kvetan V, Branka J. Assessing the representativeness of online job advertisements. Luxembourg: Publications Office. 2022. Cedefop working paper, No 17. Available from: doi: 10.2801/807500.
[6] Cedefop. Online job vacancies and skills analysis: A Cedefop pan-European approach. Luxembourg: Publications Office. 2019. Available from: doi: 10.2801/097022.
[7] Sostero M, Tolan S. Digital skills for all? From computer literacy to AI skills in online job advertisements. JRC Working Papers on Labour, Education and Technology. European Commission, Seville. 2022. JRC130291.
[8] Khurana D, Koli A, Khatter K, et al. Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications. 2023; 82: 3713-3744. Available from: doi: 10.1007/s11042-022-13428-4.
[9] Loth R, Battistelli D, Chaumartin F-R, de Mazancourt H, Minel J-L, Vinckx A. Linguistic information extraction for job ads (SIRE project). In: Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO'10). Paris: Le Centre De Hautes Etudes Internationales D'Informatique Documentaire. 2010; 222-224.
[10] Gao L, Eldin N. Employers' expectations: A probabilistic text mining model. Procedia Engineering. 2014; 85: 175-182. Available from: doi: 10.1016/j.proeng.2014.10.542.
[11] Boselli R, Cesarini M, Mercorio F, Mezzanzanica M. Classifying online job advertisements through machine learning. Future Generation Computer Systems. 2018; 86: 319-328. Available from: doi: 10.1016/j.future.2018.03.035.
[12] Cedefop European Commission, ETF, ILO, OECD, UNESCO. Perspectives on policy and practice: Tapping into the potential of big data for skills policy. Luxembourg: Publications Office. 2021. Available from: doi: 10.2801/25160.
[13] Khaouja I, Mezzour G, Carley KM, Kassou I. Building a soft skill taxonomy from job openings. Social Network Analysis and Mining. 2019; 9(1): 43. Available from: doi: 10.1007/s13278-019-0583-9.
[14] Plaimauer C. Using vacancy mining for validating & supplementing labour market taxonomies. Semantics Conference, Vienna. 2018.
[15] Ketamo H, Moisio M, Passi-Rauste A, Alamäki A. Mapping the future curriculum: Adopting artificial intelligence and analytics in forecasting competence needs. Proceedings of the10th European Conference on Intangibles and Intellectual Capital ECIIC 2019. Chieti-Pescara, Italy. 2019.
[16] Bernhard J, Russmann U. Digitalization in public relations – Changing competences: A longitudinal analysis of skills required in PR job ads. Public Relations Review. 2023; 49(1): 102283. Available from: doi: 10.1016/j.pubrev.2022.102283. doi: 10.1145/3106426.3109035.
[17] OECD. Skills for the digital transition: assessing recent trends using big data. 2022. Available from: doi: 10.1787/38c36777-en.
[18] Grüger J, Schneider G. Automated analysis of job requirements for computer scientists in online job advertisements proceedings of the 15th international conference on web information systems and technologies. 2019.
[19] Marrara S, Pasi G, Viviani M, Cesarini M, Mercorio F, Mezzanzanica M, Pappagallo M. A language modelling approach for discovering novel labour market occupations from the web. In Proceedings of WI '17. Leipzig, Germany. 2017; 1026-1034. Available from: doi: 10.1145/3106426.3109035.
[20] Lovaglio PG, Cesarini M, Mercorio F, Mezzanzanica M. Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2018; 11(2): 78-91. Available from: doi: 10.1002/sam.11372.
[21] Gurcan F, Cagiltay NE. Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. IEEE Access. 2019; 7: 82541-82552. Available from: doi: 10.1109/ACCESS.2019.2924075.
[22] Acemoglu D, Autor D, Hazell J, Restrepo P. AI and jobs: Evidence from online vacancies. Nber Working Paper Series. 2020; 28257. Available from: https//www.nber.org/papers/.
[23] Anton E, Behne A, Teuteberg F. The humans behind artificial intelligence – An operationalisation of ai competencies. In Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference. 2020. Available from: https//aisel.aisnet.org/ecis2020_rp/141.
[24] Lightcast. The digital transformation of European Labour markets. Webinar Presentation on 22 June 2023 by Durman A and Magrini E. Available from: https//lightcast.io/resources/webinars/digital-transformation-of-european-labour-markets-webinar.
[25] Chen L, Chen X, Wu S, Yang Y, Chang M, Zhu H. The Future of ChatGPT-enabled Labor Market: A Preliminary Study. 2023. Available from: arXiv2304.09823.
[26] OECD. OECD Policy Responses to Coronavirus (COVID-19) An assessment of the impact of COVID-19 on job and skills demand using online job vacancy data. 2021. Available from: https//www.oecd.org/coronavirus/policy-responses/an-assessment-of-the-impact-of-covid-19-on-job-and-skills-demand- using-online-job-vacancy-data-20fff09e/.
[27] Alipour J-V, Langer C, O'Kane L. Is Working from Home Here to Stay? A Look at 35 Million Job Ads. CESifo Forum. 2021; 22(06): 44-46.
[28] Vasilescu B. Academic papers using Stack Overflow data. [Internet]. 2012. Available from: http://meta.Stack Overflow.com/q/134495/185480 (http://meta.Stack Overflow.com/q/134495/185480) (accessed 2012).
[29] Borgonovi F, et al. Emerging trends in AI skill demand across 14 OECD countries. OECD Artificial Intelligence Papers No. 2. Paris: OECD Publishing. 2023. Available from: doi: 10.1787/7c691b9a-en.
[30] List of articles using Stack Overflow data: http://meta. Stack Overflow.com/q/134495/185480.
[31] Napierala J. The feasibility of using online job advertisements in analysing unmet EU demand. Cedefop Working Paper No 18, Luxembourg: Publications Office. 2023. Available from: doi: 10.2801/10233.