Applying machine learning tools on web vacancies for labour market and skill analysis

Emilio Colombo*

Università Cattolica del Sacro Cuore and CRISP

Fabio Mercorio†

University Milano-Bicocca and CRISP

Mario Mezzanzanica‡

University Milano-Bicocca and CRISP

February 2018, preliminary version

Abstract

This paper develops a new set of tools for labor market intelligence by applying machine learning techniques to web vacancies on the Italian labor market. Our approach allows to calculate, for each occupation, the different types of skills required. Those skills are mapped into a standard classification system. We subsequently develop measures of the relevance of soft and hard skills and in the latter group, we analyze in detail digital skills. We show that social and digital skills are related with the probability of automation of a given occupation. We also develop measures of the variation over time in the terminology used in describing occupations and finally we provide a tool for detecting new and emerging occupations through the analysis of web vacancies.

Keywords: Machine Learning, web vacancies, new occupations

JEL codes: J24, J63, C81.

*Corresponding author, Email: emilio.colombo@unicatt.it.

†Email: fabio.mercorio@unimib.it

‡Email: mario.mezzanzanica@unimib.it

1 Introduction

It is indisputable that in the past few decades significant forces and factors have dramatically changed the nature and characteristics of the labour market in both advanced and developing countries. Technical progress, globalisation and the reorganisation of the production process with outsourcing and offshoring have radically altered the demand for certain skills and competences.[1] In addition, population ageing in advanced economies intensifies the need for continued training, and is likely to affect the structural demand for certain skills, in particular those related to the health and care of the elderly.[2] The overall impact of these factors on the labor market is multifaceted. On the one hand several jobs are disappearing while new jobs are emerging; of these some are simply a variant of existing jobs, others are genuinely new jobs that were inexistent until few years ago. On the other hand the quantity and quality of the demand for skills and qualifications associated to the new labor market has changed dramatically. New skills are needed not only to perform new jobs but also the skill requirements of existing jobs have changed considerably. Which occupations will grow in the future and where? What skills will be demanded the most in the next years? Those are the questions that are at the forefront of the policy debate both among economists and policymakers. In order to address these questions specific data need to be collected. This calls for new tools for measuring and analyzing labor market trends and movements. Existing instruments are in fact either non-existent or inappropriate for measuring the complexity and the variability of new labor market trends. In this paper we develop a new set of tools for labor market intelligence by applying machine learning techniques to web vacancies on the Italian labor market. In particular those tools are specifically designed for analyzing firms skill needs. Our approach allows to shed light on a number of issues. First we can calculate, for each occupation, the different types of skills required. Furthermore we are able to classify those skills into a standard classification system and develop measures of the relevance of digital skills and of soft-hard skills. Second we show that soft and digital skills are related with the probability of automation of a given occupation. Third we develop measures of the variation over time in the terminology used in describing occupations. Finally we provide a tool

[1] See Bhagwati and Panagariya (2004); Feenstra (1998); Acemoglu (1998, 2002); Autor, Katz, and Krueger (1998); Autor, Levy, and Murnane (2003), Card and DiNardo (2002)

[2] see Freeman (2006) De Grip and Van Loo (2002).

for detecting new and emerging occupations through the analysis of web vacancies.

The remainder of the paper is structured as follows. Section 2 analyses the advantages and limits of using web vacancies with respect to other more traditional methods. Section 3 describes the methodology used, section 4 presents the results. Finally section 5 concludes.

2 Online vacancy analysis: strengths and limitations

It is natural to compare online vacancy analysis with other existing tools for assessing firms occupation and skill needs, in particular with skill surveys, which so far have been the principal tool in this area. The comparison reveals that these two approaches are very different, and in some domains at polar extremes. First the approach is opposite. Skill surveys follow a top-down approach. They have to be designed first, and the type of information collected necessarily follows from the initial design. Regarding skills, there are specific questions about them, and the list of skills is generally pre-defined. Tools based on online vacancies on the contrary follow a bottom-up approach that is entirely data-driven. The initial data collected contains all the information that individual firms post on the web. This large amount of data is subsequently filtered and processed using appropriate technical instruments to obtain the required information. In this way the tools help to categorise a pre-existing information set, but they do not pre-classify the information itself. The type of skills to be classified are those that emerge from the data, not those pre-defined in a questionnaire. This is particularly useful for the identification of soft skills and certain occupation-specific skills that surveys often ignore. The direct consequence is that with skill surveys, once the questionnaire has been created the information set is determined and not modifiable, and therefore can only be used to answer pre-defined questions identified during the survey design phase. While a Big Data approach allow one to use a new paradigm of analysis: "let the data speak", that means extract information that allow us to raise new questions and consequently to expand continuously the spectrum of knowledge of the observed phenomena. This feature is particularly important for the detection of emerging skills, as it is possible to go back to previous data in order to re-assess them. There is another direct consequence of the differences in approach: skill surveys can be designed in order to be representative of a certain population (sectors/occupations etc.). The representa-

tiveness is mostly a problem that can be dealt with by proper design, and often the limit is simply a matter of costs. Differently, representativeness of web based tools is a clear issue. It is well known that some occupations and sectors are not present in web advertisements, with the consequence that the original data set is not entirely representative. The lack of representativeness is probably the major limitation of online data.

The second key difference is related to the speed and frequency of implementation. Skill surveys are cumbersome instruments that take often months to be executed. As they are typically implemented through CATI interviews, they also involve a considerable burden of time for firms (indeed the major part of the cost of skill surveys is the opportunity cost of time for respondents). This has implications for the frequency of skill surveys, which is rarely higher than annual. Conversely, web based tools have almost no implementation lag; information collection does not need the involvement of firms or entrepreneurs. Moreover, the tools are automatically implemented by machines that can operate at any time or on any date, allowing information to be collected almost in real time.

The third major difference is related to the overall amount of the costs and their distribution along the implementation phase. Skill surveys are very costly, particularly considering the opportunity costs for the respondents. On the contrary web based tools tend to have a higher fixed cost but are subject to large economies of scale.

Another issue that has to be taken into account is the time dimension. Over the last years the number of vacancies posted on specialized and general portals has increased exponentially reflecting a more and more widespread use of the web as a relevant source for posting job offers. We expect this trend to continue so that in few years the vast majority of vacancies will be posted on the web. Indeed there is ample empirical evidence showing that online job search is crowding out alternative search channels Kroft and Pope (2014), allowing more and more individuals and firms to use web tools for searching for jobs or posting them.

Overall online vacancy analysis is a very promising approach for addressing some of the most relevant questions that new labor market trends are posing such as job specific skills, job requirements of specific occupation, identification of emerging skills in certain sectors.

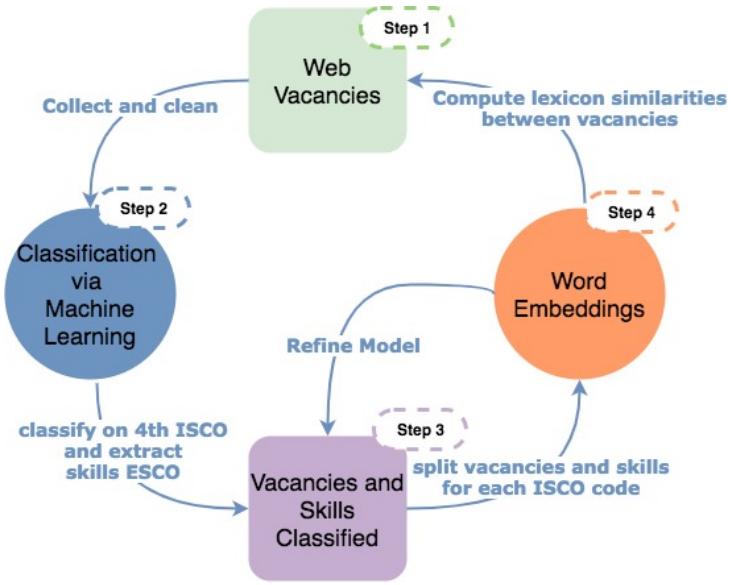


Figure 1: Overview of the framework proposed

3 Data and methodology

3.1 Data

We have used the dataset from Wollybi[3] a project that analyses online vacancies in Italy. Wollybi is now a well-established tool that has been collecting data from online job-portals since February 2013. Overall the project has developed a data base of more than 2.5 millions unique vacancies from which the tools described in the next section allowed to extract interesting informations such as location, sector, education, skills etc.[4]

3.2 Methods

From a methodological point of view, we follow the KDD approach (Knowledge Discovery in Databases, see (Fayyad, Piatetsky-Shapiro, and Smyth, 1996)), a framework that has been defined as a baseline to extract useful and reliable knowledge from raw data in real-life scenarios. It requires to apply a number of steps that we shortly summarize in Figure 1 and that can be specified as follows.

[3] See `www.wollybi.com`

[4] The project developed with Wollybi has evolved into a large and more ambitious project funded by the European Agency Cedefop that aims at establishing a real time LMI tool for the entire EU. Project Real-time Labour Market information on Skill Requirements: Setting up the EU system for online vacancy analysis AO/DSL/VKVET-GRUSSO/Real-time LMI 2/009/16.

Source Selection and Cleaning. First, data sources have be selected and evaluated. Each web-site (that includes Newspaper Websites, Employment Agencies, and Job boards) presents its own data structure, and this requires to consider all these heterogeneous data sources as a whole to build a unified data repository. To this end, each web source has been evaluated and ranked on the basis of identified criteria (typology, size, presence of the major relevant variables, quality of the information, update time, etc.). Once the sources have been identified, the data are scraped and stored accordingly. Subsequently the data have to be transformed (from one data structure to the desired one) and cleaned. Roughly speaking, transformation allows modifying the data structure and the content from the original structure to the desired/final one. During this process, the quality of the data is assessed and some cleansing activities are executed to guarantee the reliability of the data. We have applied AI algorithms for cleansing the data, as quality issues might have unpredictable effects on the analytics derived from data (see, e.g. Hernández and Stolfo (1998); Mezzanzanica et al. (2015); Boselli et al. (2014)). In our context, this task deals mainly with the identification of duplicated job vacancies posted on different Web source as well as job vacancies published multiple times (on the same site).

Classification via Machine Learning. The next step is text classification, that works by mapping web job vacancies into an existing classification system (ISCO in our case). Basically, this task requires to build a classifier: a function that maps (i.e. classifies) a data item into one of several predefined classes. In our case the items are web job vacancies whilst the classes are the ones from the ISCO 4th level hierarchy. This task has been implemented through machine learning algorithms that have been employed, and then trained on a training set of web job vacancies. Several algorithms have been applied and assessed; SVM turned out to be the best in terms of classification accuracy (Boselli et al., 2017b,a) which is higher than 93%.

More specifically text categorization aims at assigning a Boolean value to each pair $(d_j, c_i) \in D \times C$ where D is a set of documents and C a set of predefined categories. A *true* value assigned to (d_j, c_i) indicates document d_j to be set under the category c_i, while a false value indicates d_j cannot be assigned under c_i. In our LMI scenario, we consider a set of job vacancies $\mathcal{J}$ as a collection of documents each of which has to be assigned to one (and only one) ISCO occupation code. We can model this problem as a text classification problem, relying on the definition of Se-

bastiani (2002). Formally speaking, let $\mathcal{J} = \{J_1, \ldots, J_n\}$ be a set of job vacancies, the classification of $\mathcal{J}$ under the ESCO classification system consists of |O| independent problems of classifying each job vacancy $J \in \mathcal{J}$ under a given ESCO occupation code o_i for $i = 1, \ldots, |O|$. Then, a *classifier* is a function $\psi : \mathcal{J} \times O \to \{0, 1\}$ that approximates an unknown target function $\dot{\psi} : \mathcal{J} \times O \to \{0, 1\}$. Clearly, as we deal with a single-label classifier, $\forall j \in \mathcal{J}$ the following constraint must hold: $\sum_{o \in O} \psi(j, o) = 1$.

Skill Extraction. The next step involves the extraction of skills required in each vacancy through the analysis of its text. This goal is achieved by incrementally building a taxonomy of extracted words recognized as potential skill. Specifically, the system uses the n-gram[5] Document Frequency (DF), i.e., the number of vacancies where the n-gram is found. The result is a list of n-grams that identify skills together with their synonymous, we call potential skill. We then use string similarity functions[6] to link potential skills to ESCO skills and use experts to validate the linkage. Skills not linked to ESCO are useful for identifying new potential skills.

Compute Word similarities through word-embeddings. The main idea of this step is to compute the language that characterizes each occupation code through the lexicon used within the vacancies. To better understand the matter, let us to intuitively describe how word-embedding works.

Word Embedding. Vector representation of words belongs to the family of neural language models (Bengio et al., 2003) where every word of the lexicon is mapped to a unique vector in the corresponding N-dimensional space. In our context, each word used for the vacancy lexicon (that can be a term or a skill) was replaced by a corresponding vector of a multi-dimensional space.

We rely on the Word2Vec algorithm (Mikolov et al., 2013a,b) that computes the vector representations of words by looking at the context where these words are used. For example, given a word w and its context k (m words nearby w), the context k can be used as a feature for predicting the word w. The last problem can be viewed as a machine learning problem where the representation of m context words is fed into a neural network trained to predict the representation of w, according to

[5] An n-gram is a contiguous sequence of n items from a given sequence of text or speech

[6] Levenshtein distance, Jaccard similarity, and the Srensen-Dice index have been employed in this phase.

the Continuous Bag of Words (CBOW) model proposed by Mikolov et al. (2013a) [7]. Consider two different words w_1 and w_2, which have very similar *contexts*, k_1 and k_2 (e.g., synonyms are likely to have similar contexts although not equal ones). A neural network builds an internal (abstract) representations of the input data in each internal network layer. If the two output words have similar input contexts (namely, k_1 and k_2) then, the neural network is motivated to learn similar internal representations for the output words w_1 and w_2. The Word2Vec vectors are by-products of the just introduced training process; i.e., they are the neural network internal representations of the words used in the training process. For more details, see Mikolov et al. (2013b).

After the Word2vec training on the lexicon, words with similar meaning are mapped to a similar position in the vector space. For example, "powerful" and "strong" are close to each other, whereas "powerful" and "Paris" are farther away. The word vector differences also carry meaning. For example, the word vectors can be used to answer analogy questions using simple vector algebra: "King" - "man" + "woman" ≈ "Queen" (Mikolov, Yih, and Zweig, 2013).

As one might note, this approach allows representing a specific word in the N-dimensional space, while our task is to compute the vector space of *documents* (i.e., job vacancies), rather than words. We therefore apply the Doc2Vec approach (Le and Mikolov, 2014), an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents as well. As a consequence, a vector is now the N-dimensional representation of documents.

This, in turn, allows us computing the *similarity* between job vacancies in terms of both *titles* and *skills* content. The process is explained in Figure 2 that reports the pseudo-code of the algorithm we used to compute the similarity between vacancies. The algorithm takes as input (1) a set of vacancy titles $\mathcal{J} = \{j_1, \ldots, j_n\}$ along with (2) a set of skills a set of vacancies $\mathcal{S} = \{s_1, \ldots, s_n\}$, (3) the classification system $\mathcal{O}$ used to specify occupations (i.e., ISCO or SOC in this case) and the classifier that assigns a job vacancy j to one (and only one) occupation code o of the classification system used. Lines 1-7 simply initialize the local variables. Then, lines 8-12 collect job titles

[7] A similar (but reversed problem) is the Skip-n-gram model i.e., to train a neural network to predict the representation of n context words from the representation of w. The Skip-n-gram approach can be summarised as "predicting the context given a word" while the CBOW, in a nutshell, is "predicting the word given a context".

and skills for a given occupation code and a given year (e.g., *software developers* for year the 2017). Lines 13-14 apply a pre-processing pipeline for removing punctuation, special characters, stop words and to stem terms. Then, the Doc2Vec learning algorithm is applied for each set of job titles (line 15) and skills (line 16) respectively. Finally, lines 19-25 compute the cosine distance between each pair of word vector (see, e.g. Singhal et al. (2001), for details). In our settings, we aim at comparing the dissimilarity between the lexicon used in both titles and skills for a given occupation code o in a given year y. A list of dissimilarities is then returned at line 26.[8]

4 Results

4.1 Hard, soft skills and occupations

The methodology described in section 3 allows to extract skills from vacancies and map them into the ESCO system. This results in a taxonomy of skills which are associated with each occupation. The subsequent step is to group skills into categories that allow a better representation and analysis. The first major distinction is between hard and soft skills. Hard skills are typically job-specific skills and competences that are needed to perform a specific job or task (examples are knowledge of specific software or instruments, specific manual abilities etc.) Soft skills, on the other hand, are more transversal in nature and refer to the capacity of individuals to interact with others and the environment (examples are communication skills, problem solving etc.). Within hard skills we further distinguish between digital skills and non digital skills.

Vacancy descriptions are often very rich and mention several skills and competences. Given the large number of vacancies by occupation we are able to calculate, for each 4 digit occupation, the skill degree, that is the frequency of occurrence of each category or group of skill within the occupation. Figure 3 shows box plots of the distribution of the soft skill degree by 1 digit ISCO group. The figure reveals some interesting patterns. First, despite the great deal of heterogeneity of the distribution of soft skill degree within group, soft skills are pervasive and for several occupations they are more relevant than hard ones. Second, soft skills tend to be less important in low skill occupations (groups 7-9) than in high skill ones (groups

[8]The algorithm has been implemented through Python using the Gensim project libraries (Řehůřek and Sojka, 2010) for computing both Word2Vec and Doc2Vec introduced above.

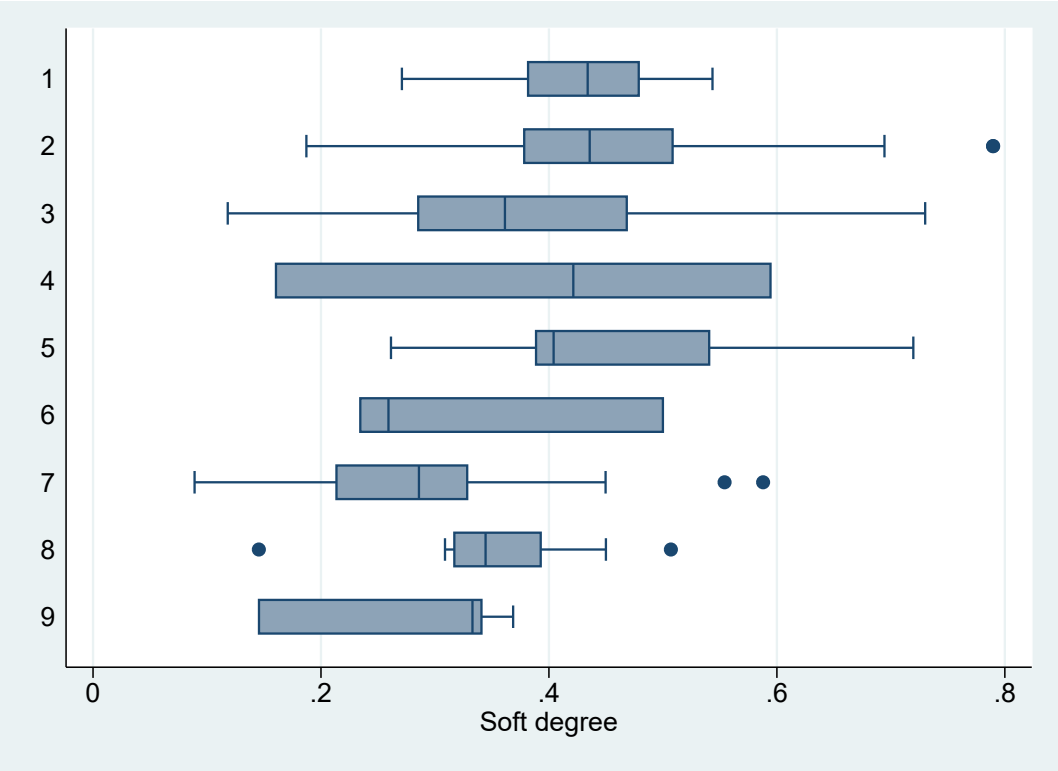
Figure 2: Pseudo-code used for computing the dissimilarities of lexicons used in web job vacancies

Data: Set of Jov Vacancy $\mathcal{J} = \{j_1, \ldots, j_n\}$
set of skills $\mathcal{S} = \{s_1, \ldots, s_m\}$
Occupation codes $\mathcal{O} = \{o_1, \ldots, o_l\}$
A classifier $\psi : \mathcal{J} \times \mathcal{O} \Rightarrow \{0, 1\}$ as described above
Result: a list $\mathcal{D}$ a list of dissimilarity values for a given o in a given year y

for $o \in \mathcal{O}$ **do**
  for $y \in year(\mathcal{J})$ **do**
    $titles_y^o \leftarrow \emptyset$ // titles of posts classified on code o in year y
    $skills_y^o \leftarrow \emptyset$ // skills of posts classified on code o in year y
    $\mathcal{M}_y^{titles,o} \leftarrow \emptyset$ // word-embedding to be computed for titles of posts classified on code o in year y
    $\mathcal{M}_y^{skills,o} \leftarrow \emptyset$ // word-embedding to be computed for skills of posts classified on code o in year y
  end
  // for each vacancy classified on code o in year y
  for $y \in year(\mathcal{J})$ **do**
    for $j \in \mathcal{J}$ *s.t.* $\psi(j, o) = 1 \wedge year(j) = y$ **do**
      $titles_y^o \leftarrow titles_y^o \cup titles(j)$
      $skills_y^o \leftarrow skills_y^o \cup skills(j)$
    end
    preprocessing($titles_y^o$)
    preprocessing($skills_y^o$)
    // Compute word-embedding for both titles and skills using Doc2Vec
    $\mathcal{M}_y^{titles,o} \leftarrow Doc2Vec(titles_y^o)$
    $\mathcal{M}_y^{skills,o} \leftarrow Doc2Vec(skills_y^o)$
  end
end
for $o \in \mathcal{O}$ **do**
  for $y_1, y_2 \in year(\mathcal{J})$ **do**
    // Use consine similarities between N-dim vectors to compute distance between two different vector spaces of words
    $\mathcal{D}_{y1,y2}^{titles,o} \leftarrow cosine\_distance(\mathcal{M}_{y_1}^{titles,o}, \mathcal{M}_{y_2}^{titles,o})$
    $\mathcal{D}_{y1,y2}^{skills,o} \leftarrow cosine\_distance(\mathcal{M}_{y_1}^{skills,o}, \mathcal{M}_{y_2}^{skills,o})$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_{y1,y2}^{titles,o} \cup \mathcal{D}_{y1,y2}^{skills,o}$
  end
end
return $\mathcal{D}$

1-3). Third, in high skill occupations and in low skill occupations the distribution of the soft skill degree tend to be more concentrated than in medium skill occupations. This could be the result of a narrower set of competences and tasks at the extremes of the distribution of occupations by skills but also by the fact that medium skill

Figure 3: Distribution of soft skill degree by Isco 1 digit group



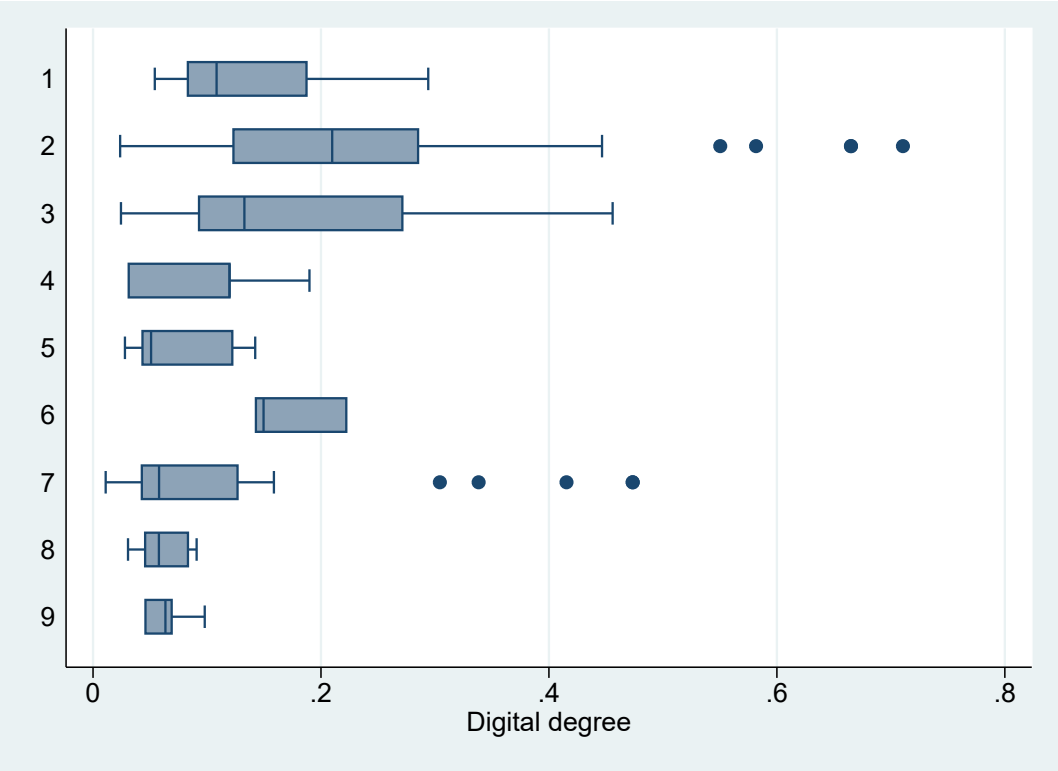
occupations are mostly affected by technological progress and globalization (Goos, Manning, and Salomons, 2014) that determine a larger demand for social skills to cope with this change. Figure 4 shows box plots of the distribution of the digital skill degree by occupation groups. Overall digital skills are less pervasive than soft skills and tend to be more relevant in high skill occupations. The next section explore more thoroughly the role of digital skills and offers a further decomposition.

4.2 Technology and jobs

In a famous study Frey and Osborne (2017) estimate the probability of computerization for a large number of detailed occupations in the US. The study spurred a considerable debate about the impact of the new technologies on the labor market.[9] Frey and Osborne (2017) identify the risk of computerization on the basis of the characteristics of selected occupations. These characteristics are derived from the O*NET system and pertain three major domains that the a group of experts identified as the major bottlenecks that computers and artificial intelligence face in completely automating a job. These domains are perception manipulation (man-

[9]For a more conservative estimate see Arntz, Gregory, and Zierahn (2016).

Figure 4: Distribution of digital skill degree by Isco 1 digit group



ual dexterity, finger dexterity, cramped workspace), creative intelligence (originality, fine arts) and social intelligence (social perceptiveness, negotiation, persuasion, care and assistance for others). These were assessed on the basis of the O*NET description of occupations. The limit of this approach is that it restricts the information domain to what is available from the official classification description. Using web vacancies it is possible to add the depth and variety of the information set of individual vacancies. In order to grasp the value added of this approach we have used the dataset of Italian web vacancies and matched the occupations with those identified by Frey and Osborne (2017). This procedure is not without problems as it is well known that there is not a one to one correspondence between SOC and ISCO classifications (see (Hoffmann, 2003)). We were able to find a match for 512 out of 702 detailed occupations in the Frey and Osborne (2017) study. For each of those occupations we have analysed the skills required and calculated the soft skill degree, the hard skill degree and, in the latter group, the ICT degree.[10]

We then used these information to enhance the information content of the classification obtained by Frey and Osborne (2017). First we have considered only the

[10]Whenever a single SOC occupation corresponded with multiple ISCO occupations we calculated the soft, hard and digital degrees by averaging across ISCO occupations.

occupations that have been classified by experts in FO's study[11] for which we have explained the probability of automatisation on the basis of the hard/soft/digital skill degree. The first two columns of Table 1 report the results of the probit analysis and shows that the probability of automation is positively correlated with the hard skill degree and negatively with the soft skill degree. This is in line with the FO approach in fact hard skills are more technical and more likely to be replaced by machines or software whereas soft skills are less automatable. Interestingly the digital degree is never significant in explaining occupation automability. This is explainable with the fact that digital and ICT was not a feature considered by experts when classifying occupations as automatable. Columns 3-7 extend the analysis to the full sample of 512 occupations. In this case the dependent variable is the degree of automability as estimated by FO in their study. We implement both simple OLS and weighted OLS where we weight occupation observations on the basis of the frequency of vacancies in our sample. The results confirm the negative correlation between soft skills and automability; interestingly now the degree of digital skill is statistically significant and negatively related with the probability of automation. This show the potential advantage of using information on skills. FO classified occupation automability on the basis of job characteristics. Adding information about skills allows to understand where skills can temper the negative impact of technology on jobs. Digital skills are exemplary since they can complement the use of machines and software and therefore make the job less substitutable even for occupations that are on average highly automatable. Table 2 develops further the interactions between hard and digital skills by dividing the sample at different values of the distribution of hard skill degree. For simplicity, given the limited number of observations we have split the sample in two: above and below the median[12] At low levels of hard skill degree hard skills tend to be negatively correlated with the probability of automation, this correlation disappears when we consider occupations characterized by a high hard skill degree. However digital skills are always negatively correlated with the probability of automation irrespective of the degree of hard skills. Columns 3-4 perform a similar analysis considering different values of the distribution of the probability of automation. For occupations with low probability of automation hard skills are positively related with the probability of automation while digital skills are negatively correlated. Considering occupations characterized by higher probability of

[11] We have an exact match only for 44 of such occupations

[12] Similar results are obtained using quartiles of the distribution.

automation the hard skill degree is now not significant while the digital degree now turns positively related with the probability of automation. This can be explained with the fact that occupations characterized by the highest probability of automation are often medium level administrative occupations (clerks, tellers, credit analysts) for which are required basic digital skills such as the use of spreadsheet that tend to be substituted by technological advances rather than complement them. In order to investigate this issue further we have further refined the textual analysis and split the classification of digital skills into four subgroups as described below.

Information Brokerage Skills. Refer to the ability to use ICT tools and platforms for data exchange and communication (e.g. social media);

Basic Informatics Skills. Refer to the ability to use some ICT specific applications for supporting the individual professional activities (e.g. use of spreadsheet or word processing software);

Applied/Management Informatics Skills. These skills refer to tools and software used within the organisation for supporting management, operational and decision making processes (e.g. administrative software);

ICT Technical Skills. Refer to solutions, platforms and programming languages that are strongly related to ICT-specific professions (e.g. programming languages, advanced ICT softwares).

Columns 5-6 of Table 2 report the analysis using this decomposition of digital skills outlined above. Confirming our intuition for highly automatable occupations the probability of automation is positively correlated with basic digital skills. On the contrary more sophisticated ICT technical skills are always negatively correlated with the probability of automation.

Table 1: Explaining the probability of automation: soft vs hard skills

	Rest. sample Probit	Rest. Sample Probit	Full sample OLS	Full sample W. OLS	Full sample OLS	Full sample W. OLS
Soft skills	-7.047***		-0.727***	-0.645***		
	(2.047)		(0.122)	(0.001)		
Hard skills		7.476***			0.842***	0.714***
		(2.445)			(0.118)	(0.001)
Digital skills		-0.598			-0.719***	-0.683***
		(1.735)			(0.110)	(0.001)
Const.	3.020***	-4.145***	0.817***	0.759***	0.169**	0.224***
	(0.827)	(1.336)	(0.045)	(0.000)	(0.079)	(0.001)
R2	0.283	0.281	0.065	0.058	0.137	0.148
N	44	44	512	512	512	512

Note: Cols 1-2 dep. variable = 1 if occupation is automatable 0 if not. Cols 3-6 dep. variable = probability of automation. W. OLS = weighted regression, weights are the number of vacancies per occupation. Cols 1-2 report pseudo R2. * denotes significance at 0.05 level, ** at 0.01.

Table 2: Probability of automation and Digital skills

	Q12 Hard	Q34 Hard	Q12 Prob	Q34 Prob	Q12 Prob	Q34 Prob
Hard skills	1.365***	-0.415	0.589***	-0.008	0.573***	0.016
	(0.273)	(0.300)	(0.106)	(0.048)	(0.105)	(0.049)
Digital skills	-0.977***	-0.779***	-0.493***	0.101**		
	(0.261)	(0.113)	(0.099)	(0.044)		
Information Brokerage					-0.048	0.015
					(0.381)	(0.148)
ICT Technical					-0.694***	-0.190*
					(0.179)	(0.106)
Basic Information					-1.264***	0.383***
					(0.349)	(0.139)
Applied Management ICT					-0.249	0.052
					(0.183)	(0.093)
Const.	-0.076	1.137***	-0.004	0.868***	0.025	0.853***
	(0.139)	(0.232)	(0.064)	(0.034)	(0.067)	(0.036)
R2	0.101	0.158	0.143	0.021	0.182	0.068
N	256	256	256	256	251	249

Note: Q12-Q34 Hard define respectively quartiles 1,2 and 3,4 of the distribution of Hard skill degree. Q12-Q34 Prob refer to the distribution of the probability of automation. * denotes significance at 0.05 level, ** at 0.01,

4.3 How does the vacancy contents of occupation change?

As stressed in the introduction the major factors that are affecting the labor market are responsible for the change of the skills and qualifications needed for occupations. The identification of the change in the skill content of occupations is one of the most challenging problem in labour market analysis. To deal with this issue we have applied the word embedding tools described in section 3 by measuring the differences in lexicon within occupations (4 digit ISCO) between 2014 and 2017 in our sample. In other words for each occupation we have analysed whether and to what extent the content of the vacancy changed e.g. how the description changed. We measured the change along two main dimensions. First we have analysed the evolution of the distribution of job titles, second the evolution of the distribution of required skills. Each vacancy contains a job title which is generally expressed in natural language, i.e. not coded according to a standard classification. The machine learning tool described in section 3 classifies each vacancy in the standard ISCO system. This means that for every occupation we have the distribution of job titles that are used to describe it. Similarly for the text we can obtain the entire distribution of words used for skills in the description of the vacancies that refer to the same occupation.

In Figure 5 we report four box-plots, one for each classification algorithm. Each box-plot shows the distribution of the dissimilarity distribution value (left side plots) and the lexicon variation (right side plots) for each ISCO first-level groups.

The left-hand side of Figure 5 shows the variation of the vocabulary by comparing terms used in job titles (upper-left plot) and skills extracted from vacancies (lower-left plot) posted in 2014 and in 2017 for a given ISCO code, for identical sources scraped. This metric highlights the lexical richness of the Labor Market Information from Web. As one might note, the vocabulary size has considerably grown in terms of job titles used to advertise jobs and skills requested since 2014. In other words the terminology used to identify occupations and to describe the associated skill set considerably increased over a short period of time. It is difficult however to infer from this a substantial change in the required skills driven by the megatrends identified in the introduction. In fact the increase in the vocabulary could be simply the result of an increase in the number of synonyms used that does not correspond to a substantial change in the skills required for a specific occupation. In order to investigate this further we have analyzed the variation of the *lexicon* used, using the

Doc2Vec techniques as described above. This solves the problem above since the algorithm recognizes word synonyms and skills used in similar contexts. For example the algorithm would recognize that a term such as *software developer* used in 2014 is related to a more recent term as *app developer* by learning the context of words that characterized the lexicon used in these vacancies. As expected lexicon variation is considerably smaller than vocabulary variations. Furthermore, the median value of the lexicon variations for skills is higher than the one computed on titles only, showing that there seems to be more dynamism in terms used to express skills, rather than for terms used for advertising job positions. The size of the variation is also interesting as it shows that on average there is a 35-40% variation in the lexicon used to describe skills for groups of occupations. Finally a closer inspection of the median values shows that there is an interesting pattern emerging in the distribution of skills (bottom right panel). The largest lexicon variation for skills refers to the top and the bottom of the skill distribution (groups 1-3 and 7-9). This confirms the findings of the literature that stresses that technological change is determining a polarization in the labour market of advanced economies (Michaels, Natraj, and Reenen, 2014); as a consequence high and low skill occupations are the ones where skill lexicon changes the most.

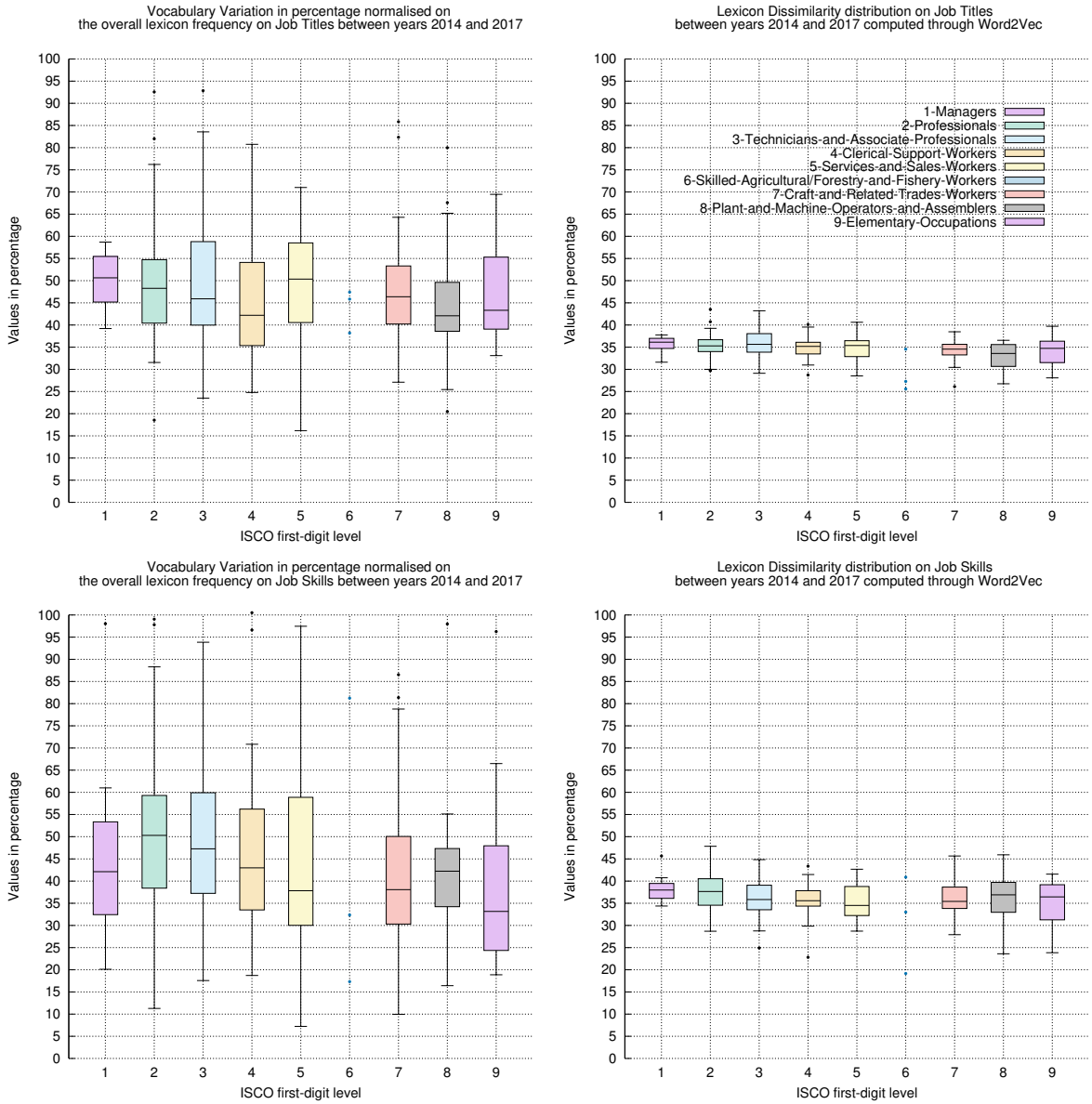
4.4 New emerging occupations

One of the most debated issues in the literature that studies the effect of technological change on the labour market is whether and to what extent the new technologies not only destroy or modify existing jobs but also create new ones. But how to detect new jobs? The issue is gaining importance due to the increasing speed of technological progress. The typical approach in detecting new and emerging occupations is to exploit changes in the official classification. For instance Lin (2011) identifies new jobs by collecting new occupation titles from U.S. classification indexes in 1977, 1991, and 2000.[13] This procedure however allows to detect new occupations only ex post, i.e. after a change in the classification which generally occurs after a certain number of years.

During the las year the Bureau of Labour Statistics (Occupational Employment Statistics, 1998) and subsequently the O*NET system for O*NET Development (2006)

[13] More specifically Lin (2011) exploits information from the Dictionary of Occupation Titles and from the Census Bureaus Classified Index of Industries and Occupations

Figure 5: Vocabulary Variation and Lexicon Dissimilarity on Job Titles between years 2014 and 2017.



established a methodology for the identification and classification of new occupations. This methodology defines as "new" an occupation, not adequately described by the existing classification, which "involves significantly different work than that performed by job incumbents of other occupations, as determined by NC State and O*NET research consultants".

This methodology has several merits but has the limit of involving a strong component of expert judgment in the initial identification of the critical industries where new occupations should be detected. The expert judgment is necessary in order to limit the amount of qualitative analysis undertaken, given the extent and variety of the impact of technological progress on the labour market.

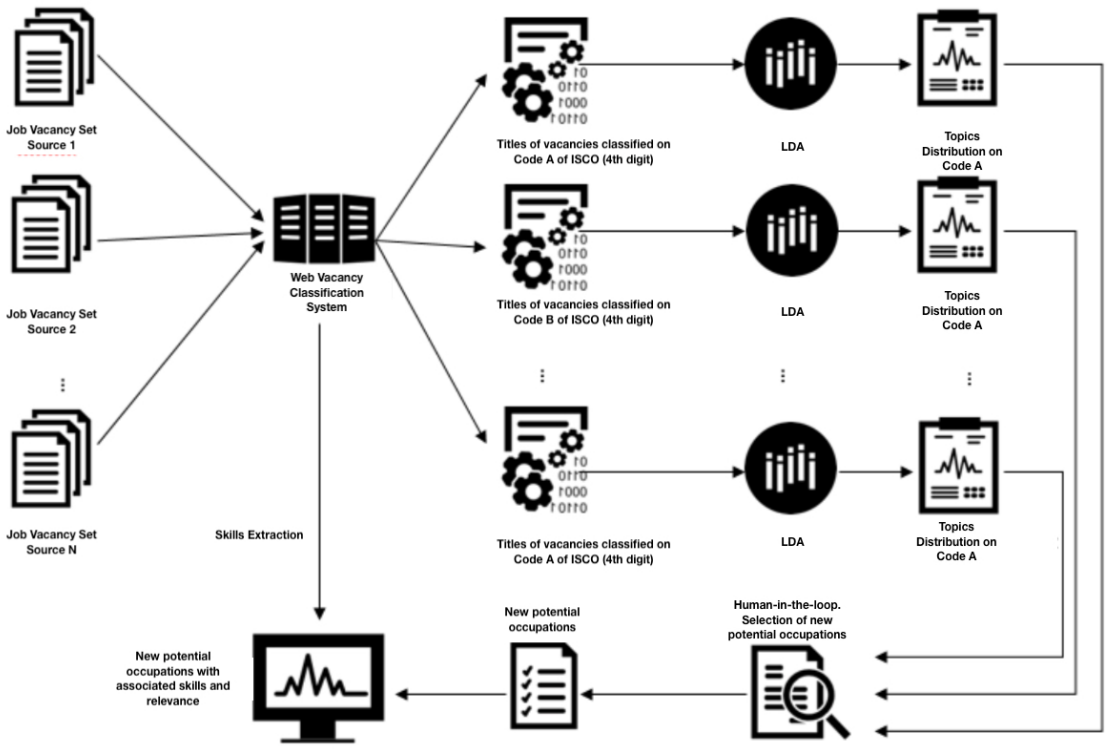
The analysis of web vacancies allows to develop a new approach to the identification of new and emerging occupations. This approach is more data driven and exploits all the richness of web vacancies but also constructs some quantitative metric.

A Human-in-the-loop approach for detecting new emerging occupations. In the following we describe how topic modeling can be used to identify new (potential) emerging occupation, by exploiting a unsupervised learning techniques. More specifically, let us suppose to have a collection of documents - composed of no more than 10 or 15 words each - whose content is a mixture of arguments (e.g. *topics*, $T_1, T_2, \ldots, T_n$) that characterize the lexicon of each subset of documents. LDA (Latent Dirichlet Allocation Blei, Ng, and Jordan (2003)) is a generative probabilistic model that considers each document as a mixture of latent topics, where each topic is characterized by its own words distribution. LDA allows clustering documents by topics on the basis of the frequency of words. As a result, each topic is composed by words that mostly contribute to its generation. The higher the probability that a topic contains a certain term, the higher the relevance of that term has in the corresponding topic. LDA does not make a partition of terms by topics, as a term can belong to more than one topic, albeit having a different relevance.

The idea behind the use of LDA for identifying new potential occupations relies on considering a job vacancy's title as a document whose content might be ideally composed of a number (fixed but unknown) of topics to be identified. Figure 6 provides a graphical overview of how this process works. Once each job vacancy has been classified on the standard taxonomy (i.e., ISCO 4th digit in our case), the LDA

algorithm is applied on each subset of vacancies, grouping them by ISCO codes. This pre-selection phase would help LDA on reducing the features space and to maximize the LDA performances as well. The LDA process returns a number of topics along with their probability distribution of words (n-grams) that compose each topic.[14] The process returns a list of top-terms for each ISCO code, that has to be analyzed and refined by a labor market specialist. Since LDA is an unsupervised learning, a human supervision that validates the final outcome is mandatory to guarantee a reliable result. Finally, terms that identify new potential occupations are linked to job vacancies in which that terms have been found, and then linked to the skills included. This allows one to compute the new emerging occupations and to filter-out only *skills* requested by them. A selection of some new occupations found using this approach is provided in Tables 3 and 4, that also show the hard and soft skills found respectively. The tables show that, as expected, most of the new occupations are in the ICT sector and require very technical skills. However also for technical occupations social skills are extremely important due to the increasing pervasiveness of ICT in very different domains.

Figure 6: Human-in-the-loop approach for discovering new potential occupations



[14] Notice the the number of topics to be identified is an input parameter of any LDA-based approach that has to be tuned properly.

Table 3: New emerging occupations selected along with the corresponding top Hard-skills. The upper text of each dashed line is the top level skill in the ESCO skill hierarchy (v0) with its relevance within the occupation. The lower text of each dashed line reports the most important skills extracted and belonging to the parent ESCO skill. (*) Represents skills extracted but not present yet in the ESCO taxonomy (i.e., new skills).

New Occupation	**Hard Skill**	**Hard Skill**	**Hard Skill**	**Hard Skill**	**Hard Skill**
Brand Manager	*Informatics (54.83%)*	*Business & Administration (43.24%)*	*Mathematics & Statistics (1.93%)*		
	MS Office SAP CRM Management Software*	Public Relations Marketing Management	Data Analysis		
Business Analyst	*Informatics (47.62%)*	*Business & Administration (26.19%)*	*Mathematics & Statistics (25.54%)*	*Law (0.75%)*	
	MS Office SAP CRM & SQL Google Analytics & ERP*	Public Relations Management Customer Relationship Management	Data Analysis Data Analysis	Legal Studies Security Law	

Table 3 Continued

BI Analyst	*Informatics (50.32%)*	*Mathematics & Statistics (29.60%)*	*Business & Administration (19.33%)*	*Law (0.75%)*
	SQL & Oracle SharePoint & Data Integration*	Data Analysis	Public Relations Management Manage Quality	Legal Studies Security Law
Data Scientist	*Mathematics & Statistics (48.15%)*	*Informatics (29.01%)*	*Business & Administration (22.84%)*	
	Data Analysis SAS SAP & SPSS*	SQL & Java Business Intelligence Data Integration*	Public Relations Management Customer Relationship Management	
Data Analyst	*Informatics (46%)*	*Mathematics & Statistics (39.02%)*	*Business & Administration (13.70%)*	*Law (1.29%)*
	Business Intelligence	SAS	Management	Legal Studies

Table 3 Continued

	SQL & Java	Data Analysis	Public Relations	Security Law	
Facility Manager	*Business & Administration (48.07%)*	*Informatics (19.89%)*	*Law (16.02%)*	*Architecture & Constructions (8.28%)*	*Mathematics & Statistics (7.74%)*
	Management	MS Office	Security Law	Electrical Installations	Data Analysis
	Public Relations	AutoCAD	Legal Studies	Electrical Installations	
	Negotiation experiences	Basics of Informatics		Reading of Technical Draws	
HSE Specialist	*Law (41.51%)*	*Informatics (28.30%)*	*Business & Administration (16.98%)*	*Mathematics & Statistics (5.66%)*	*Physics (3.77%)*
	Security Law	Basics of Informatics	Manage Quality	Data Analysis	Mechanics
	Legal Studies	SAP CRM	Industry Systems		
Regulatory Affairs	*Business & Administration (36.41%)*	*Law (31.79%)*	*Informatics (28.72%)*	*Mathematics & Statistics (3.08%)*	
	Public Relations	Legal Studies	MS Office	Data Analysis	
	Management	Security Law	Basics of Informatics		

Table 3 Continued

	Customer Relationship Management	Security Law	SAP CRM

Table 4: New emerging occupations selected along with the corresponding top Soft-skills

New Occupation	Soft Skill	Soft Skill	Soft Skill	Soft Skill	Soft Skill
Brand Manager	Foreign languages (45.34%)	Positive attitude (31.71%)	Cooperation with others (7.69%)	Leadership ability (6.61%)	Problem solving (6.21%)
Business Analyst	Foreign languages (46.52%)	Positive attitude (30.83%)	Problem solving (10.99%)	Cooperation with others (5.47%)	Leadership ability (2.20%)
BI Analyst	Foreign languages (39.23%)	Positive attitude (31.51%)	Problem solving (15.35%)	Cooperation with others (6.27%)	Leadership ability (4.34%)
Data Scientist	Positive attitude (38.58%)	Foreign languages (29.63%)	Problem solving (13.58%)	Cooperation with others (6.79%)	Leadership ability (4.63%)
Data Analyst	Positive attitude (40.91%)	Foreign languages (34.85%)	Problem solving (8.71%)	Cooperation with others (5.30%)	Leadership ability (5.30%)
Facility Manager	Foreign languages (39.68%)	Positive attitude (28.95%)	Leadership ability (16.09%)	Problem solving (11.53%)	Values (2.95%)
HSE Specialist	Foreign languages	Problem solving	Positive attitude	Cooperation with others	Leadership ability

Table 4: New emerging occupations selected along with the corresponding top Soft-skills

New Occupation	**Soft Skill**	**Soft Skill**	**Soft Skill**	**Soft Skill**	**Soft Skill**
	(14.55%)	(56.36%)	(20.00%)	(3.64%)	(1.82%)
Regulatory Affairs	Foreign languages	Positive attitude	Problem solving	Leadership ability	Cooperation with others
	(54.19%)	(26.32%)	(6.10%)	(5.14%)	(3.71%)

5 Conclusions

In this article we develop a new set of tools for labor market intelligence by applying machine learning techniques to web vacancies on the Italian labor market. These tools are specifically designed for analyzing firms skill needs. Our approach allows to shed light on a number of issues. We can calculate, for each occupation, the different types of skills required. We are able to classify those skills into a standard classification system and develop measures of the relevance of soft and hard skills in the latter group, we can drill down detailed digital skills. We show that social and digital skills are related with the probability of automation of a given occupation. We also develop measures of the variation over time in the terminology used in describing occupations and finally we provide a tool for detecting new and emerging occupations through the analysis of web vacancies. Overall this approach provides extremely promising insights that allow to better understand the relevant changes that are affecting the labor market.

References

Acemoglu, Daron. 1998. "Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality." *The Quarterly Journal of Economics* 113 (4):1055–1089.

———. 2002. "Technical Change, Inequality, and the Labor Market." *Journal of Economic Literature* 40 (1):7–72.

Arntz, Melanie, Terry Gregory, and Ulrich Zierahn. 2016. "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis." Social, employment and migration working papers, no. 189,, OECD.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger. 1998. "Computing Inequality: Have Computers Changed the Labor Market?" *The Quarterly Journal of Economics* 113 (4):1169–1213.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics* 118 (4):1279–1333.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. "A neural probabilistic language model." *Journal of machine learning research* 3 (Feb):1137–1155.

Bhagwati, Jagdish and Arvind Panagariya. 2004. "The Muddles over Outsourcing." *Journal of Economic Perspectives* 18 (4):93–114.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3 (Jan):993–1022.

Boselli, Roberto, Mirko Cesarini, Stefania Marrara, Fabio Mercorio, Mario Mezzanzanica, Gabriella Pasi, and Marco Viviani. 2017a. "WoLMIS: a labor market intelligence system for classifying web job vacancies." *Journal of Intelligent Information Systems* .

Boselli, Roberto, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. 2014. "Planning meets Data Cleansing." In *The 24th International Conference on Automated Planning and Scheduling (ICAPS)*. 439–443.

———. 2017b. "Using Machine Learning for Labour Market Intelligence." In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part III*, *Lecture Notes in Computer Science*, vol. 10536. Springer, 330–342.

Card, David and John E. DiNardo. 2002. "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles." *Journal of Labor Economics* 20 (4):733–783.

De Grip, Andries and Jasper Van Loo. 2002. "The economics of skills obsolescence: A review,." In *The Economics of Skills Obsolescence*, edited by Andries de Grip, Jasper van Loo, and Ken Mayhew, Research in Labor Economics, Volume 21. Emerald Group Publishing, 1–26.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. "The KDD Process for Extracting Useful Knowledge from Volumes of Data." *Commun. ACM* 39 (11):27–34.

Feenstra, Robert C. 1998. "Integration of Trade and Disintegration of Production in the Global Economy." *Journal of Economic Perspectives* 12 (4):31–50.

for O*NET Development, National Center. 2006. "New and Emerging (N&E) Occupation. Methodology Development Report." Tech. rep., O*NET.

Freeman, Richard B. 2006. "Is A Great Labor Shortage Coming? Replacement Demand in the Global Economy." NBER Working Papers 12541, National Bureau of Economic Research, Inc.

Frey, Carl Benedikt and Michael A. Osborne. 2017. "The future of employment: How susceptible are jobs to computerisation?" *Technological Forecasting and Social Change* 114:254 – 280.

Goos, Maarten, Alan Manning, and Anna Salomons. 2014. "Explaining Job Polarization: Routine-Biased Technological Change and Offshoring." *American Economic Review* 104 (8):2509–2526.

Hernández, Mauricio A and Salvatore J Stolfo. 1998. "Real-world data is dirty: Data cleansing and the merge/purge problem." *Data mining and knowledge discovery* 2 (1):9–37.

Hoffmann, Eivind. 2003. "International Statistical Comparisons of Occupational and Social Structures. I." In *Advances in Cross-National Comparison*, edited by J.H.P. Hoffmeyer-Zlotnik and C. Wolf. Springer.

Kroft, Kory and Devin G. Pope. 2014. "Does Online Search Crowd Out Traditional Search and Improve Matching Efficiency? Evidence from Craigslist." *Journal of Labor Economics* 32 (2):259–303.

Le, Quoc and Tomas Mikolov. 2014. "Distributed representations of sentences and documents." In *International Conference on Machine Learning*. 1188–1196.

Lin, Jeffrey. 2011. "Technological Adaptation, Cities, and New Work." *The Review of Economics and Statistics* 93 (2):554–574.

Mezzanzanica, Mario, Roberto Boselli, Mirko Cesarini, and Fabio Mercorio. 2015. "A model-based evaluation of Data quality activities in KDD." *Information Processing & Management* 51 (2):144–166.

Michaels, Guy, Ashwini Natraj, and John Van Reenen. 2014. "Has ICT Polarized Skill Demand? Evidence from Eleven Countries over Twenty-Five Years." *The Review of Economics and Statistics* 96 (1):60–77.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* .

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*. 3111–3119.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." In *Hlt-naacl*, vol. 13. 746–751.

Occupational Employment Statistics. 1998. "New and emerging occupations in occupational employment and wages." Bulletin 2506, Bureau Labour Statistics.

Řehůřek, Radim and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 45–50.

Sebastiani, Fabrizio. 2002. "Machine learning in automated text categorization." *ACM Computing Surveys* 34:1–47.

Singhal, Amit et al. 2001. "Modern information retrieval: A brief overview." *IEEE Data Eng. Bull.* 24 (4):35–43.