DataOps for Societal Intelligence: a Data Pipeline for Labor Market Skills Extraction and Matching

Damian A. Tamburri  
Technical University Eindhoven  
Jheronimus Academy of Data Science  
s’Hertogenbosch, The Netherlands  
d.a.tamburri@tue.nl

Willem-Jan Van Den Heuvel  
Tilburg University  
Jheronimus Academy of Data Science  
s’Hertogenbosch, The Netherlands  
w.j.a.m.v.d.heuvel@jads.nl

Martin Garriga  
Tilburg University  
Jheronimus Academy of Data Science (JADS)  
s’Hertogenbosch, The Netherlands  
m.garriga@uvt.nl

Abstract—Big Data analytics supported by AI algorithms enable skills localization and retrieval, in the context of a labor market intelligence problem. We formulate and solve this problem through specific DataOps models, blending data sources from administrative and technical partners in several countries into cooperation, creating shared knowledge to support policy and decision-making. We then focus on the critical task of skills extraction from resumes and vacancies featuring state-of-the-art machine learning models. We showcase preliminary results with applied machine learning on real data from the employment agencies of the Netherlands and the Flemish region in Belgium. The final goal is to match these skills to standard ontologies of skills, jobs and occupations.

I. INTRODUCTION

The ability of our society at large to harness the digital ecosystem and turn it into an opportunity is known as societal intelligence. In particular, job seekers and employment agencies alike could leverage Big Data analytics supported by DataOps methods and AI algorithms, to turn the act of finding the right skills into a labor market intelligence problem. The challenges we face in this context are manifold: (i) Integration of data sources from administrative and technical partners (job seekers, employers, trainers, public authorities, etc) in several countries; (ii) Extracting available/required skills and employment options, especially in low-compliance areas (e.g., borderlands); (iii) Matching skills from resumes with vacancies automatically, leveraging structured information about Skills, Competences and Occupations from standard ontologies and taxonomies; and finally (iv) Assess, Improve and Extend such ontologies and taxonomies with the descriptions of novel, arising job profiles and skills.

In this context, the main contributions of this paper target the following research questions:

**RQ1.** How to model a DataOps Pipeline to enable Societal and Labor Market Intelligence in the context of skills extraction and matching? We design such pipeline addressing the aforementioned challenges, to move from Big Data to Knowledge. Then we illustrate how to automate and scale the analysis of job vacancies and resumes, and then matching against representative ontologies (such as ESCO/ISCO).

**RQ2.** How to extract the skills from vacancies and resumes in an effective and efficient way? We instantiate a pipeline with hundreds of thousands vacancies stemming from the Dutch-Flemish labor market and then solve the skills extraction problem featuring state-of-the-art machine learning models.

The rest of this paper is structured as follows. Section II presents background definitions. Section III describes the DataOps design for Societal and Labor Market Intelligence, and Section IV how to instantiate it for the skills extraction and matching problem. Section V presents the validation with real data of the Dutch-Flemish cross-border labor market. Section VI discusses main findings and limitations. Finally, Section VII concludes the paper.

II. STATE OF THE ART

**a) Societal and Labor Market Intelligence:** Societal intelligence is the family of approaches, services, and platforms that adopt Big data analytics, machine-learning or other technologies specific to business intelligence over social, organisational, and societal data for the purpose of a more instrumented, self-sustainable society. Furthermore, Labor market intelligence (LMI) refers to the design and use of Big Data and AI algorithms and frameworks to analyze labor market information for supporting policy and decision-making [2], [3], [4]. Labor market information encompasses skills, competencies, qualifications, and occupations, including the ICT techniques and services to manage such information, in particular, mobility-related services. Key to LMI is the identification and adoption of standard taxonomies for occupations and skills, to foster the circulation of information in a multi-language job market like the European one.

**b) Labor Market Ontologies:** The standard ontology to classify occupations for the international labor market is ISCO [5] (International Standard Classification of Occupations). ISCO is one of the most adopted taxonomies in Europe and it is a reference worldwide, while the United States developed its own, namely O*NET1.

For skills description one can find broad regional classifications such as the multilingual taxonomy ESCO (European Skills, Competences and Occupations2). ESCO extends ISCO with fine-grained occupation descriptions, and a taxonomy of skills, competencies and qualifications.

Specific classifications can be found for each country or cross-border region. As a relevant example within this work, Competent3 is a taxonomy developed in Belgium and recently

---

1https://www.onetonline.org/
2https://ec.europa.eu/esco/resources/data/static/model/html/model.xhtml
3https://competent.vdab.be/competent/
extended to the Netherlands as Competent-NL\footnote{http://production.competent.be/competent-nl/main.html}. Competent and other fine-grained classifications can be easily linked to ISCO and ESCO through mapping tables [6].

c) DataOps: Like DevOps, DataOps aims to combine the production, operation and delivery (of data) into a single, agile practice that directly supports specific business functions to improve quality, speed, and collaboration and promote a culture of continuous improvement [7]. A DataOps methodology combines and interconnects data engineering, data integration, data quality, and data security/privacy [8] to deliver data from its source to the person, system, or application that can turn it into business value [9].

III. BIG DATA FOR SOCIETAL INTELLIGENCE: A PROOF-OF-CONCEPT FOR THE LABOR MARKET

Knowledge extraction from (Big) labor market data has been addressed using the knowledge discovery in databases (KDD) approach [10] – as shown in Figure 1, with the span of the Vs of Big Data throughout the process.

a) Step 1: Select Data Sources: Each Web source is evaluated to guarantee its reliability – e.g., vacancy publication date, update frequency, (un)structured data, privacy restrictions – including but not limited to, governmental employment offices (from different regions/countries), private Websites (e.g., indeed.com) and social networks (e.g., LinkedIn).

b) Step 2 Pre-processing Selected Data: This includes data cleansing tasks to remove noise or inappropriate outliers (if any), deciding how to handle missing data, as well as removing duplicated entries. Data quality and cleaning tasks are indeed mandatory steps of any data-driven decision-making approach for guaranteeing the trust on the overall process [3].

c) Step 3: Transformation: This brings data into a unified model, which depends on the goal of the process, by means of data reduction and projection techniques in the context of an ETL (Extract, Transform, Load) pipeline [11]. Around 70% of the effort is spent on Data Selection, Pre-processing and Transformation, to provide input data in the right amount, structure and format for each Data Mining task [12].

A crucial transformation for Societal and labor Market Intelligence consists of extracting the important information – i.e., the skills (hard and soft) in the job vacancy text. More formally, the transformation function takes a job vacancy \( j_k \in J \) as input and extracts a set of skills \( S_k = \{s_1, \ldots, s_n\} \) where \( S_k \subseteq P(S) \), the powerset of skills \( S \). Note that the skills set \( S \) evolves throughout time as novel skills arise.

d) Step 4: Data Mining and Machine Learning: This step involves text classification algorithms to map vacancies/resumes into one of several predefined classes in a taxonomy or ontology of job profiles and skills, such as ISCO, ESCO and/or Competent. More formally, categorization assigns a Boolean value to each pair \((j_k, c_i) \in J \times C\) – where \( J \) is job vacancies and \( C \) the predefined categories (of a given taxonomy) – \textit{true} if \( j_k \) belongs to \( c_i \) and \textit{false} otherwise. This categorization task can be solved through machine learning: let \( J = J_1, \ldots, J_n \) be a set of job vacancies and the classification of \( J \) under the taxonomy consists of \( [O] \) independent problems of classifying each job vacancy \( J_i \in J \) under a given taxonomy occupation code \( o_i \) for \( i = 1 \ldots [O] \). Then, a classifier is a function \( \phi : J \times O \rightarrow \{0, 1\} \).

e) Step 5: Interpretation and Feedback: Finally, this step employs visual paradigms to represent the resulting knowledge, according to the ultimate goal. For example, public administrations might be interested in identifying the most requested occupations in a cross-border area; companies might focus on monitoring skills trend and novel, arising skills and occupations, so that they can design training paths.

IV. A DATAOPS PROCESSING SOLUTION FOR SKILLS EXTRACTION AND MATCHING

The critical task of skills extraction and matching entails a DataOps pipeline combining Big Data and Machine Learning models to process vacancies and resumes, and extract the sentences containing skills. Figure 2 shows Model Training and Prediction pipelines from our prototype, detailed below.

a) Data Pre-processing: Given the characteristics of our input (vacancies text), several text preprocessing steps are necessary, mainly comprising (but not limited to): (i) Distinguish end-of-sentence periods (.) from others (e.g. abbreviations, acronyms, websites, etc.); (ii) Distinguish end-of-sentence line breaks (\(\backslash n\)) from spurious ones; (iii) Trim leading/trailing spacing/tabbing; (iv) Identify and replace special characters used as bullet points (such as hyphens, dashes or even ‘o’); and (v) Filter stopwords with little lexical meaning, such as pronouns and articles. The output of this step is a list of sentences ready to be used by the Machine Learning algorithm.

b) Sentences Annotation for Learning: Each sentence in the training set must be annotated by domain experts: the ones that contain a skill are annotated with a 1, otherwise with a 0. In the case of ambiguous sentences, we created a codebook with help of the domain experts as an aid to identify and correctly annotate the sentences, with previously agreed definitions on what is (not) a skill.

c) Model Training: First, the model is trained on unlabeled data over different pre-training tasks. Pre-trained models can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications [14].

d) Supervised Training (fine-tuning): Then a training set (‘train’ data) consisting of the labeled instances is given to form a description that can be used to predict unseen examples [12]. This is pre-processed (as described above), tokenized and converted into the format that the model understands – typically csv or tsv files, containing the columns with the annotations, the vacancies/resumes and skills ids for traceability, and the actual sentence text.

At this point the model can be actually trained for the needed purposes. Early stopping with a configurable number of epochs is recommended, as the training theoretically should finish when the model achieves the optimal state under the
setup conditions, avoiding overfitting. Afterwards, the accuracy can be verified using the ‘dev’ validation dataset, a labeled set not directly used for training. The output of this process is the trained model which can be used for further predictions.

Note that, compared to pre-training, fine-tuning is relatively inexpensive. All of the results in the following section can be replicated in at most 1 hour on a single Cloud TPU, or a few hours on a GPU, starting from the same pre-trained model.

e) Predictions with the Pre-trained ML Model: At this point the model is ready to make predictions, according to the pipeline depicted in Figure 2. Recall that the goal of the pipeline is to classify (predict) sentences describing skills from a vacancy text. Pre-processing is equal to the one performed for training. Then the sentences extracted from the processed vacancies/resumes text (in the proper tsv/csv format) are given as input to the model, pre-trained and fine-tuned during the training pipeline. The output contains the predictions for each sample, with the original text, ids and class probabilities (whether the sentence contains a skill or not).

V. VALIDATION: THE CASE ON THE CROSS-BORDER DUTCH-FLEMISH LABOR MARKET

We applied the proposed DataOps pipeline in the context of project Werkinzicht5, which aims to provide a clear picture of the Dutch-Flemish cross-border labor market for job seekers, employers/agencies, educational institutions and governments. Cross-border job boards are available from UWV and VDAB – Dutch and Flemish employment agencies respectively.

We used three alternatives for the underlying Machine Learning Model in Figure 2. First, a Simple Neural Network6 as baseline. Then, two alternative deployments of the best model to date for this task: BERT (Bidirectional Encoder Representations from Transformers) [14], namely: BERT in the Cloud7, running on Google Cloud with a single TPU device with 4 TPU chips and 8 cores; and BERT Local deployment8 running on on-premise hardware.

The input dataset was provided by UWV – consisting of millions of vacancies from year 2014 onwards. For the training and validation activities, the domain experts curated a total of 10K vacancies that became the gold-standard. After pre-processing it we obtained a total of 300K sentences.

a) Annotation: Six domain experts from UWV annotated a total of 3K sentences – around 500 sentences per person.

Each sentence is annotated with a ‘1’ if contains a skill, and as ‘0’ if it does not. The annotation process was aided by the Codebook9, describing what is intended as a Skill, and how to annotate the dataset. We performed two rounds of annotations plus feedback, and also cross-checked the annotations made by the experts with the ones made independently by the authors, to reach unanimity on the skills definition and identification. The final result for the Inter-rater reliability assessment through the Kappa [15] coefficient (that measures the agreement) amounted to 0.86, higher than the typical reference score of 0.80 (i.e., 80% agreement).

b) Model Training: We trained the models according to the steps discussed in Section IV. The baseline pre-trained model used both for BERT Cloud and BERT Local options is the BERT Multilingual Cased Model10. From the 3K annotated sentences we adopted a 75/25 splitting for training (train set) and validation (dev set) respectively.

c) Results: The accuracy results are shown in Figure 3. The best performing model in this case is BERT Cloud, with an accuracy of 0.816 after 210 training steps (epochs), followed by BERT Local (0.8) and Simple-NN (0.78). All three models show high accuracy, although it could be improved by means of training with more annotated sentences, fine-tuning hyperparameters, etc., as future work. Table I shows the confusion matrix for the best performing model – i.e., BERT Cloud. Note the high recall (0.94) of annotated skills correctly predicted. Moreover, as stated by the experts from UWV, the worst case scenario are false negatives, i.e., skills that are incorrectly classified as not skills. The model has a very low false negative rate of 0.05.

VI. DISCUSSION AND LIMITATIONS

In this section we recap on the RQs posed in Section I. To answer RQ1 on the DataOps pipeline for Labor Market Intelligence, Section III presented our design to move from big data to knowledge, providing societal and labor administrators

---

5An EU Interreg project – https://werkinzicht.eu/
6Source and documentation: https://tinyurl.com/y86rerhw
7Source and documentation: https://tinyurl.com/ybytgj52
8Source and documentation: https://tinyurl.com/ycks8sp4
9https://tinyurl.com/8yk34xqw (in dutch)
10https://github.com/google-research/bert/blob/master/multilingual.md

---

Fig. 1. Big Data enables Societal and Labor Market Intelligence. The process specifies which Vs of Big Data are involved in each step (adapted from [3])
with Big Data analytics and business intelligence. This shows the competitive advantage of automated analysis of Web job vacancies in terms of, e.g., (i) near-real-time labor market information; (ii) reduced time-to-market (iii) fine-grained cross border territorial analysis [3].

Then, to answer RQ2 regarding the skills extraction problem, we instantiated our pipeline within the prototype, featuring state-of-the-art machine learning models with hundreds of thousands vacancies stemming from the Dutch-Flemish labor market. When validating our prototype in terms of accuracy, BERT Cloud was the best performing model. It also features the fastest execution time, as it leverages on-demand infrastructure from Google Cloud, resulting in speedups of up to 10x for training and predictions—from hours/days to minutes. That should be taken into account when deploying this prototype in production, given the scalability and performance challenges.

Moving to limitations, our models have been trained on UWV data and as such only encompass Dutch vacancies and resumes. This can bring along cross-border applicability issues as e.g., Flemish and Dutch languages may differ. However, the underlying BERT multilingual models have been successfully trained in one language and then used for prediction on another [16]. Another concern is data privacy and governance: putting sensitive data to the Cloud may be an issue according to GDPR (General Data Protection Regulation). In these cases, the local deployments may overcome the cloud ones because of privacy and governance, although the tradeoffs with performance and scalability should be carefully considered.

VII. CONCLUSIONS

This article introduced the design for a DataOps intelligence analytics platform to support a more sustainable labor market. This was partially validated with real data from the Dutch prototype for the key tasks of skills extraction from vacancies and resumes. Results show both technical feasibility and high accuracy and recall. In the future we plan to gather hard data over the areas present in our case-study to refine and bring to production our prototype; and integrate the overall pipeline for skills matching against existing ontologies, to finally refine them to reflect the highly dynamic labor market.

REFERENCES

[1] E. Hoffmann, “International statistical comparisons of occupational and social structures,” in Advances in cross-national comparison. Springer, 2003, pp. 137–158.
[2] R. Boselli, M. Cesarini, S. Marrara, F. Mercorio, M. Mezzanzanica, G. Pasi, and M. Viviani, “Wolmis: a labor market intelligence system for classifying web job vacancies,” Journal of Intelligent Information Systems, vol. 51, no. 3, pp. 477–502, 2018.
[3] M. Mezzanzanica and F. Mercorio, “Big data for labour market intelligence: An introductory guide,” Eur. Training Found., Torino, Italy, Tech. Rep. 2019.
[4] UK Commission for Employment and Skills, “The importance of labor market intelligence.” UK Government, Tech. Rep., 2015. https://goo.gl/TDrwS.
[5] ILO International Labour Office, “Iesco-08 international standard classification of occupations.” International Labour Office, Tech. Rep., 2012. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---pubdocs/publication/wcms_172572.pdf.
[6] ESCO European Commission on Skills, Qualifications and Occupations, “Final report on the esco mapping pilot.” European Commission, Tech. Rep., 2016, https://tinyurl.com/ycezrh52.
[7] J. Ereth, “DataOps—towards a definition.” Proceedings of LWDA (Lernen, Wissen, Daten, Analysen) Conference, pp. 104–112, 2018.
[8] A. Palmer, “From devops to dataops,” TAMR Inc, 2015, https://www.tamr.com/blog/from-devops-to-dataops-by-andy-palmer/.
[9] J. Easton, “What is dataops?: Platform for the machine learning age,” Jul 2018. [Online]. Available: https://www.nextlm.com/define-dataops.
[10] U. Fayyad, G. Piatesky-Shapiro, and P. Smyth, “The kdd process for extracting useful knowledge from volumes of data,” Communications of the ACM, vol. 39, no. 11, pp. 27–34, 1996.
[11] P. Vassiliadis, A. Simitos, and S. Skiadopoulos, “Conceptual modeling for eTL processes,” in Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP, 2002, pp. 14–21.
[12] F. H. a. Salvador García, Julián Luengo, Data Preprocessing in Data Mining, 1st ed., ser. Intelligent Systems Reference Library 72. Springer International Publishing, 2015. [Online]. Available: http://gen.lib.rus.ec/book/index.php?md5=F91A1699CE0E23DE40B59A8F0F9797EC.
[13] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, “Yale: Rapid prototyping for complex data mining tasks,” in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 935–940.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
[15] K. Krippendorff, Content analysis: An introduction to its methodology. Sage publications, 2018.
[16] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual bert?” arXiv preprint arXiv:1906.01502, 2019.