Title of publication: "The feasibility of using Big Data in anticipating and matching skills needs”
Author/Editor: Ana Podjanin, Olga Strietska-Ilina and others
Publisher: International Labour Office
Publication year: 2020
Number of pages: 113

Overview:

The paper summarizes the findings of ILO’s experts during its 2019 conference on the use of Big Data, like online job vacancies, in identifying and matching skills demand. The main focus of the discussion will be on the implementation of skills analytics using Big Data.

- Share practical challenges from best practices in developed countries
- Examine solutions to common challenges
- Discuss how best to implement Big Data-based analytical methods in developing countries

Finally it was established that Big Data is a source of valuable data and information from which the skills can be determined and matched. Nevertheless, it was demonstrated that single analysis is not an complete answer. Big Data should be seen as a supplement to existing approaches depending on the situation in each country and intended application.

1. Introduction

■ Background and Objectives of the Study

The rate at which the labor market is changing has greatly increased due to technological advancement and global economic expansion. Demand for skills understanding has become a very crucial issue. Nevertheless, traditional survey techniques tend to be limited given problems related to costs and time delay hence calls for novel ways of getting timely information.

Recently, one method getting increasingly popular includes forecasting of real time skill demands using large amounts of data on online jobs vacanciess. This gives particular skills details but the data quality and representativeness are highlighted as well.

This paper aims at reviewing the current situation of Big Data-based skills predictions and the challenges they encounter in developed economies with emphasis on some of the best practices for their utilization. We also, will discuss the relevance of this approach to developing nations. The objective of this research is to come up with policy that will address the dynamic labour market.

■ Overview of research approach and methodology

【Research approach】

This paper seeks to explore the possible applications of Big Data in skill demand forecasting and matching. Therefore, analyze how the experts from the ILO's 2019 Expert Workshop discussed their contributions in this regard. This is how we have tried to compile the current knowledge on Big Data analytics and highlight what is challenging and possible for the future.

【Research methodology】

The study methodology encompasses the following four phases.

Phase 1: Concepts for organizing data management and technicalities.

- Analyze issues surrounding online job information and analysis methods.

Phase 2: Examples and lessons learned from using Big Data applications in developed countries.

- Using instances of Big Data implementation in the developed countries of US and Europe coupled with their repercussions.

Phase 3: Assessing feasibility and identifying constraints in developing nations.

- Assessing suitability and challenges in forthcoming economies like India, Africa and Latin America

Phase 4: Appropriate combinations of existing methods and detection of problems and prospects.

- Future study questions, benefits for such merger with conventional interviewers, and other considerations.

The above four phases suggest strategic application of the method per country situations.

2. Concept and technical aspects of Big Data

■ Definition of Big Data

Gartner defines Big Data with these three words – volume, velocity, and variety, which are now considered as the "4 V's". These four Vs include volume, variety, velocity, and value.

1. To describe it in simpler terms, Volume refers to an exceptionally large volume of data size.
2. The word "for variety" simply implies the addition of different kinds and forms of data, such as structural data as well as other non-structural data like texts, pictures, videos, and audios.
3. For Velocity, it means that data gets generated and retrieved within seconds while being analyzed simultaneously.
4. Value is what can be mined out of these Big Data in terms of discovery of new ideas and useful decisions.

Moreover, truthfulness or trustworthiness (Veracity) is another point that has to be taken into account. Accuracy in analysis can only be achieved with the veracity and reliability data collected since the large amount of data collected is mostly prone to include errors or duplicates.

Big Data analysis aims to obtain precise information and inference from enormous datasets by employing machine learning as well as other AI algorithms in real time.

Big Data like online job posting is known for its volume, speed as well as diversity of available formats. However, with proper pre-processing and analytic techniques, much meaningful information can be generated herein from the large volumes of real time job data to infer changes in labor demand and supply. As such, the internet job listing data is presented as an information resource which can be used in examining the Big Data approach.

■ Collection and Analysis Process

The analysis of labor market information using online job vacancies is conducted primarily through the following steps.

1. First, an appropriate website from which to get online job information should be chosen. Demand trends vary from one job site to another. Therefore, it is necessary to collect job data from different job sites for understanding overall demand trends thoroughly.
2. Subsequently, huge chunks of job textual data are extracted from selected job sites employing APIs or web scrapping and crawler. To cope with such high volumes of unstructured data, automation is necessary.
3. It is essential that pre-processing of collected data takes place. Due to its nature, data preprocessing and cleaning should involve completing of missing data, removal of duplicates, and noise removal because there are so much data for it. This ensures high integrity of data truthfulness and improves data analysis precision.
4. Automated extraction of occupational roles and competencies from job texts. For computers to process huge volumes of data that cannot be done manually, unstructured textual data should be structured and normalized. The abovementioned structuring process employs machine learning and ontology.
5. Lastly, the resultant data is converted into indicators associated with occupation, region, time series among others and the analysis report made available to relevant parties, using web dashboards and other means. Appropriate visualizations of the analysis results will be helpful as they are challenging to do a semantic interpretation on.

It is necessary for the skill-related vocabulary in the job text to be correctly extracted and defined so as to provide insights on skills from online job vacancies. Nevertheless, fuzzy skill expressions like “digital skills” and “communicative skills” will be useless for the analysis. These skill definition have to cover what constitutes certain skills and knowledge.

Thereafter, these thus-defined individual skills should then fit within an integrated skill system, and although the framework of a standard skill classification system like ESCO is a starting point for that, merely applying it is far from resolving the incompatibilities between countries or languages. Moreover, the framework needs to be country specified and tailored according to realities of every language region.

Construction of an ontology that would cover semantic relations between super ordinate concepts, sub-concepts and related concepts is even more important. It is only through such a method that it could be possible to establish the systematics and semantic connections between the skills.

High-quality analysis realization starts with this series of skill definition and semantics being seen as essential preprocessing step. This forms a strong foundation for organizing huge volumes of disorganized data.

■ Significance and Limitations of Online Job Posting

【Significance】

- Obtaining Large Amounts of Data at Low Cost
  With relatively lower costs and effort as compared to traditional surveys, large volume of job data can easily be acquired through automatic collection of online public information. Most importantly, the API linkage facilitates cheap and continuous data collection in real time.
- Continuous Acquisition of Baseline Data
  Having routine data collected enables continuous access to baseline information on periodic basis without conducting special surveys. Conducting pulse checks on the underlying trend in the labour market as a whole.
- Building a Highly Comprehensive Analysis Infrastructure
  Collecting a broad scope of job data from varied job sources can be used in forming an expansive analytically reliable database with a negligible bias. This eases generalisation of results.
- Identification of Emerging Occupations
  It is possible to detect emerging occupations at an early stage by making use of machine learning that analyses time-changes series of new occupations' names appearing in job text. However, this is achievable as the mining method for frequent pattern extraction applies to large volumes of text data.
- Understanding Skill Ratios
  Disaggregating specific hard and soft skills across regions and occupations will make it possible to analyze, in which specific contexts the different skills show up. This is due to newly developed techniques for skills extraction from big texts.
- Use for Demand Forecasting
  Online posted skill frequency data may serve as an input to forecasting future occupational or skill demand with historic series. The accuracy of forecasts is mainly dependent on the use of a lot of real data.

【Limitations】

- Discrepancy between the number of jobs and actual market demand
  In some cases, the number of jobs posted online differs from the actual market demand for jobs. This implies that an ideal picture of the total labour market does not arise.
- Low percentage of jobs in small, medium, and micro companies

The number of jobs on larger companise is very much over-represented compared with the one's put on small, medium and the micro firms. This makes it difficult to assess the impact on supporting industries.

- Difficulty in promptly detecting emerging occupations
  In some cases, it is difficult to assess the impact of emerging occupations on the labor market in a timely manner.
- Cost of processing alternative representations and cost of countermeasures against missing, duplicate, and noisy data
  There is an array of alternative terms that appear in recruitment advertisements, and their pre-processing and normalization is expensive. This cleaning and pre-processing is also costly.
- Difficulty in analysis due to bias in the amount of information posted
  Information about current job vacancies differs by employer size and job types thereby causing problems when developing homogeneity for general analytical platform.
- Unevenly distributed number of listings by region/occupation
  The differences in remote area and small to medium jobs could be responsible for an imbalanced number of job postings by job type, making it hard to have complete information about the entire labor market.

【Complementarity with existing methods】

- Cedefop methodologies such as the European Skills Forecast:
  In addition to the existing Cedefop methodologies for recognition of mid–to–long term labor market and skills trends.
  This method supplements the existing Cedefop methodology that can detect mid- and long-term changes in the labour market and skills but also identifies both cyclical short-term fluctuations and mid- and long-term changes in these domains.
- Segment analysis based on the European Classification of Industries and Occupations (NACE/ESCO):
  It can retrieve trends of a job in an occupation segment or from an industrial basis to find jobs that are posted online using a connection between the NACE and ESCO standard classification to the job postings.
- Analysis by region (NUTS):
  By applying NUTS (National Units of land Survey and Planning), which are common regional classifications in Europe, we can make comparisons between regions within a country, and conduct cross-country analysis.
- Matching with the results of the European Skills and Occupations Survey:
  Combining them with those from the previous survey will yield comparative results. This helps fill in any gaps and provides a wider understanding of the labor market when used alongside conventional survey outcomes.

It is promising especially when it comes to real-time performance as well as how fragmented the system can be. However, this creates problems with the complexity of data cleaning and analysis processes. Moreover it has a high risk of data bias and is complimented with conventional surveys. As such, using it as a credible single source is an issue right now.

3. Case Studies in Developed Countries

■ Summary of practices in developed countries (Austria, Netherlands, Canada)

【Australia】

The PES (the Public Employment Service) in Austria enlisted Textkernel to analyze employment statistics, and executed a pilot study on AI to verify as well as extend its competency categorizing technique.
The assessment by the Austrian PES, in validating the skills taxonomy, used the following data.

- Taxonomy term list consisting of over 29,000 words
- Online job text data of over 850,000 jobs
- Job seeker profile information

The results showed that as many as 56%, taxonomy terms were not used at all in actual jobs. This is a clear indication that the taxonomy used has a very significant difference with the field's language.

In addition, there was a negative correlation as the top ten most frequently repeated words were short and transversal in nature while the lengthier they became lesser did their use become. This underpinning highlights the importance of being concise when creating taxonomy term definitions as well as improving user friendliness of NLP system.

On the other hand, thanks to the project we managed to employ AI technology for auto-extracting 1,900 particular newly discovered words which would have otherwise been missed by a manual inspection. The process ended with the addition of seventy-nine new words to the taxonomy, confirming a high impact of automatic AI support. Although the approach of classification intends to be a complete one, in practice it cannot sufficiently meet all needs of modern NLP. Simplification needs to be followed closely with the application of AI for making continuous evidence-based improvements in the future.

【Netherland】

In the Netherlands, the SBB Foundation is implementing "Job Perspectives," a calculation project of the job chances for new students based on job data. The project estimates employment probability for the new students based on the various data, which include:

- Macroeconomic forecasts from the Bureau of Economic Policy Analysis
- Demographic and employment trend data from the Central Bureau of Statistics
- Scraping data from online job platform
- Online surveys of 250,000 companies

These enabled calculations of employment supply and demand trends for every industry and occupation, thereby giving employment prospect information on available qualifications to both students and educational establishments. Using a perfect combination of structural data coupled with the large-scale Big Data, we were able to comprehend the current jobs opportunities.

However, small and medium-sized enterprises have less job posts available in comparison with large companies, and there are few on-line posts made as well. The study showed that this lack of information could affect the true demand for the project. The results suggest that if there is any bias when using job data alone then additional data ought to be taken concurrently. This means that jobs in small- and medium-sized companies are not so easily posted online, hence giving us a false picture of the matter at hand. Therefore, the data must be integrated into the structural data.

【Canada】

Canada is working towards joining the classification systems that the department of employment and social devolution have employed in their competencies with the national occupational classifications. The classification is composed of five skills groupings which include fundamental, analytical, and technical; and seven holistic categories. One can comprehensively comprehend the necessary skills of the occupation by mapping this against the NOC's occupational profile. The challenges that currently need to be overcome include:

- Reduces reliance on occupational analysts, who are inefficient and prone to information obsolescence, and overcomes cost and update frequency constraints
- Reflects actual job conditions that cannot be covered by structural data
- Reflects stakeholder input in real time

To solve these problems, we plan to build a hybrid platform with immediacy, comprehensiveness, and transparency, taking advantage of the characteristics of Big Data and structural data. They thus propose a highly sensitive mapping system by using a feedback function. In particular, they strive to associate professions and knowledge with the specified information sources.

- Structured data from official statistics
- Big Data from online job vacancies
- Analysis results from occupational analysts

■ Common challenges and lessons learned

【Common Issues】

- Risk of bias in jobs and skills data and deviation from overall trends
  The judgments can be biased on the basis of the skills and jobs data, especially when no information is available concerning small and medium-sized enterprises in the labour market as a whole.
- Constraints on frequency and speed of updating structural data
  Government statistical surveys usually update slowly, with lapses before they are turned into data.
- Data linkage with structural data for which appropriate integration is essential

It is difficult to link with methods of comprehensible analysis for proper linkages with structure data.

- Inconsistency of classification system and ontology, ambiguity in skill measurement
  Classification systems of occupations, skills, etc. vary across the countries and the organizations. Skill measurement has space for subjectivity.
- Cost and labor issues
  Handling data at a large scale is usually hard, time consuming, and quite expensive.

These three cases supported an element of skewness bias in single data like jobs and competencies, plus discrepancies away from general trend. These issues included, infrequent updates, and delayed provision of structural data.

【Important Lesson Learned】

- Appropriate combination with structural data is vital and complementary is essential.
  It is necessary to combine not only Big Data but also structural data to complement each other's weak points.
- Continuous input from stakeholders is essential.
  There would be a need for incorporating feedback from continuous stakeholder engagement among others.
- Ensuring quality and stability of the entire data ecosystem is key
  To ensure the reliability of analysis results, it is necessary to guarantee the quality and stability of all data used.
- Objective skill definitions that are faithful to the data are required.
  In enhancing the reliability of analysis results, objective skill definitions derived from the data are required.
- Sustainable data infrastructure and operational structure must be established.
  It is vital for long-term operation to base on sustainable platform in structure and software side.

This approach provided evidence as to why it is crucial to employ a combined Big Data and structural data approach. Further, the study has been found to have provided evidence for the effectiveness of the feedback driven PDCA cycles used as the core of the system in order to sustain the overall data quality.

4. Case Studies in Developing Countries

■ LinkedIn's Potential Use for Big Data Analysis

LinkedIn is one of the most popular professional social networks with more than 330 million users worldwide and exceptional information about the demand-side of labour market. It is distinct from other such platforms, most notably because its members disclose extensive information such as employment history, abilities, and training. People put very specific data concerning their employment record, professional competences, qualifications, etc.

1. The profile information is updated in real time. The flexibility of the labor market as it responds to technological changes is constantly measured by looking at member’s records that are being updated and revised frequently since they constantly show changes in the workplace or in job skills. It has this characteristic that sets it apart from ordinary data like statistical surveys.
2. Detailed indicators of skill importance can be obtained for each occupation and country. For example, it can be used to compare and analyze the skill sets of software engineers by country. This can be used to develop human resource development strategies that are tailored to regional differences.
3. A huge amount of data allows one to provide qualitative analyses of labor market tendencies. Researchers could use this dataset, among other things, as a point of reference for the development of evidence-informed policies aimed at measuring current growth occupations and transition potential between jobs/occupations.

Such characteristics make LinkedIn data a credible information basis for policy making and skill development programs. This includes identification of the mismatch between supply and demand and analysis the possibility of transition between occupations. Nevertheless, it was observed that since this piece of information is self-provided by members, it does not always correspond with real ability needs. Likewise, the sample was based on “digital-native” individual.

■ Summary of Cases in Emerging Countries (Latin America, India, Myanmar)

【Latin America and the Caribbean】

Analysis of LinkedIn data reveals that demand for technology-related skills such as software development and digital marketing has grown significantly. On the other hand, demand for basic digital skills has stagnated among managers and other positions that traditionally required these skills.

LinkedIn users are a skewed population that includes mostly higher-educated members hence one problem to data application. The study applied data in order to facilitate a smooth job shift that was from declining professions to growing ones. These include developing of tools which visualizes the possibility of moving among occupations and reforming of education programmes.

【India】

The economic graph enabled a clear view of the movement in supply and demand of the artificial intelligence related professions like the software engineers and data scientist. Gender bias was also manifested in analysis of entrepreneurial data. However, tackling hidden structural problems continues to be a difficulty.

However, it should be noted that this analysis could be useful in predicting the skill demand required and identifying the movement of the labour force. Measures to correct the gender imbalance as well as enhance inclusiveness serve as examples of what would be required in future endeavours.

【Myanmar】

Analysis of data from online job sites provided insights that complemented traditional labor statistics by visualizing occupations and skills in high demand. Although there are some limitations in the scope of data and classification system as challenges, the data is highly useful in directing skill development.

It is imperative in future to develop a sophisticated cross-skilled governance, bringing together Big Data analytics with traditional statistical systems.

■ Differences from Developed Countries, Common Challenges

【Differences】

A bias in LinkedIn talent data is also higher in emerging economies than developed countries because they mostly have a lag in digital skills diffusion. For instance, in Latin America, it focuses on people who are highly educated.

In addition, the reliability of labor data and the standardization of occupational classifications are not sufficiently developed, so there are significant limitations when analyzing the data.

【Common Challenges】

Like developed countries, new economies encounter a similar problem that necessitates adjustment to dwindling birth rates, ageing populations, and automatization or digitalization. For instance, data in India show a shortage of personnel specialized in AI.

Moreover, it necessitates putting up a shield between education and jobs involving changes in cultural milieu at workplaces and those that cannot be extracted out of available statistics. This calls for collaboration between the public and private sectors.

5. Combination with Other Data Sources

Various attempts are currently being made to combine Big Data with other data sources. The reasons for this are discussed below.

Skill demand analysis using Big Data (and especially online job data) has proven to be an effective and valid strategy. This can include examples of limitations like low internet penetration in developing countries.

However, traditional survey data also carry their own limitations. Regular surveys are difficult to conduct among developing countries due to quality assurance constraints. Scholars have highlighted the inflexibility inherent within skills mapping in even the developed nations.

Thus, the two could combine to offer a joint and credible evaluation that would be more useful. This includes for instance, complementary validation of qualitative research outcomes, filling up of information voids in underdeveloped nations and dividing data into small fragments but made real-time.

Case studies are essential for critically examining existing challenges that need to be addressed as well as ongoing efforts.

■ Summary of case studies combining various data

【Case Study 1】: ILO's "Skills for a greener future" report

This study was conducted to assess the effects of environmental policy on skills and employment as well as green skills changes for a sustainable economy. It started with a country case study and skills development survey in 32 countries on how economic activity and skills influenced policy. Next, we used EXIOBASE, a quantitative computational model capable of analyzing inter-industry trade in 44 countries, to estimate changes in demand for skills by occupation under two scenarios: toward energy sustainability and circular economy transition. Thereafter, we leveraged on related U.S. online job information to discover “transferable skills” among occupations as also particular skill adjustments required within development fields. Doing this allows a global and quantitative analysis of employment and skill re-orientation changes under the policy scenarios, which serves as a good input to make the right decisions.

【Case Study 2】: ILO's STED program

This program sought to analyze essential skills required for trade success, and formulate plans to overcome the problem of skill shortages as well as scarcity gap; extensive reviews were made on more than thirty industrial sectors and twenty states, and information were found wanting in developing nations. As such, we used U.S. job data as a proxy and estimated the changes in skills demands in four manufacturing recovery states. We were then able to bridge the information insufficiencies in various developing nations and establish the real nature of skills needed.

【Case Study 3】: Joint research between ILO and OECD

This study aimed to develop a global skill demand shortage/surplus index based on the combination of skills demand index by job vacancies data and labour force survey data. We constructed an index capable of international comparisons on basis of comparison of the supply-demand gap using two primary sources – the job data from which we identified top occupation required skills and labor force survey data from where we calculated occupational shortage index. More can also be done to increase the scope of this index, such as by incorporating data for jobs from other parts of the world. Such changes would improve the segmentation, as well as the real time performance on the index.

■ Complementary Effects and New Insights

【Complementary Effects】

In the case study 1, The qualitative country studies gave a critical assessment of the situations in the countries, prevailing policies, and collection of good practices. However, the quantitative model enabled the quantifying of effect under the policy scenario. The combination enabled the investigation from various angles such as analyzing the status quo and making predictions the future prospects. We were also able to understand more detailed changes on skills by using job data.

The deficiency of information of developing countries was supplemented by data on skilled change reflecting actual conditions from developed countries in this program for case 2. The concept of reciprocity between data has worked well.

The construction of an indicator comprising of international comparability, segmentation, and real-time performance in Case 3 through a successful combination of job offer data and labor force survey data has achieved success.

【New Insight】

The first case unearthed new findings that include skill shifts in occupations. As a result, it was evident that the environmental policy also changes the abilities needed for every profession.

In turn, Case 2 enabled us to measure a novel association between job alterations and shift in skills in the manufacturing sector. Important also was our ability to determine needed skills due to shifting of demand.

Importantly, in Case 3 we succeeded in getting an index enabling comparison of the mismatch between skill demands and labor forces on a global scale.

6. Conclusion

■ Significance and Limitations of Labor Market Information Surveys Using Big Data

【Significance】

- Complementary clarification of non-capturable changes through skills demand trends recognition from actual job data in real time./
- Given these limitations on internet penetration in developing economies, at best, missing information on labour markets could serve as an alternate means for this purpose and assist in prediction of skills needed.
- Such an analysis would be beneficial as it will relate skill requirements with occupations and qualification included in the job openings thus providing the necessary information required for formulation of policies.

【Limitations】

- The nature of the job postings may not fully reflect the actual situation because the skill requirements for low-skilled occupations can only be fully understood.
- The number of job postings in developing countries is limited because of the lack of Internet access, so there is a limit to the amount of information that can be used as a substitute.
- Due to privacy protection considerations, the possibility of using Big Data for administrative purposes remains an issue. Analysis requires expertise and ethical considerations.
- It cannot be a direct substitute for periodic statistics such as the Labor Force Survey, and can only be positioned as a complement.

- **Future Research Directions**
  - At present, job related data predominate, so the possible expansion of other types of Big Data such as public administrative data or data from national state statisticians could become a topic of further research.
  - To make the method more applicable to developing countries they have to add more job offer data per country and make an analysis possible by using local languages.
  - There are requirements for improving, upgrading, and deepening the quality of data to keep pace with changes so as to improve real-time functionality more.
  - The combined mixture of periodic statistics and data is perceived to exhibit a greater degree of resistance towards environmental change.
  - Companies and governments should have strong data science-trained human capital for effective utilization. Therefore, proper understanding and consideration of ethical issues are going to be more crucial than ever before.

While avoiding excessive expectations at present, initiatives are spreading around the world, and it is likely to be seen as a phase of steady development while understanding the significance and limitations of Big Data.

Potential Use Cases for Labor Market Information Systems with Big Data in Qatar

- Evaluating the Economic impacts of diversification policy.
  Through Big Data it is possible to track on time the emergence of a new industry for diversifying economies as well as changes in diversification indices. It is also used to gauge policy effectiveness.
- Verification and optimization of policy in the education industry.
  The Big Data analyzes the efficiency and effectiveness of educational policy influence on the labour market in order to optimize.
- The methodology of research on expanding domestic talent recruitment
  Big Data can be used to conduct a survey on domestic human resource demand and willingness to learn for effective planning of domestic human resource recruitment policy.
- Forecasting Demand for foreign talent in medicine and healthcare sectors.
  Another big question is hiring highly qualified foreign human resources in health field. Using job opening database, it is possible to forecast the niche specialization, number, and skills the international workforce will need.
- Enhancing construction planning in transport infrastructure.
  It can use up-to-date information from the construction sector of the construction industry data for flexibility in the tempo and man-hours for the transport infrastructure construction. It is anticipated that this will enable them realize their cost targets in time.