European Network on Regional Labour Market Monitoring

NOTE

Seminars of the Big Data Knowledge Hub

Exploring new sources of Labour Market Intelligence: job seekers sentiment analysis

Notes of the online seminar promoted by the Big Data Knowledge Hub of the European Network on Regional Labour Market Monitoring (ENRLMM). April 28, 2022

The first of the Seminars of the Big Data Knowledge Hub took place on April 28, 2022. The aim of these series of seminars to offer an opportunity to deepen the Network's knowledge on how to use Big Data for labour market research and consulting by presenting practical cases and demonstrations.

The study “Exploring new sources of Labour Market Intelligence: job seekers sentiment analysis" is presented by **Ciprian Panzaru** (Universitatea de Vest din Timișoara, Romania) and following comments are made by **Moreno Baruffini** (Università della Svizzera Italiana, Switzerland). The open discussion counted with the participation of: **Connie Boulandier** (Ayto. Leioa – Behargintza, Spain), **Christa Larsen** - IWAK Goethe University Frankfurt am Main, Germany), **Aleksandra Webb** (University of the West of Scotland, UK), **Claudiu Brandas** (Universitatea de Vest din Timișoara, Romania) and **Joanna Napierala** (Cedefop).

Introduction

Eugenia Atin (Big Data Working Group of the ENRLMM) and **Christa Larsen** (ENRLMM), after the initial greetings and thanks to the participants, highlight the relevance of big data as a tool for describing the regional labour market and analysing the demand side which is always more challenging than the supply side. The ENRLMM has for years been exploring new sources of information and data to describe the demand side and, in 2015, promoted by the University of Milano Bicocca who had the technological resources and means, Big Data was the topic of the year and the ENRLMM established a working group to explore the new opportunities that big data could bring to regional labour market observatories.

We realised then that some sources of big data are available and accessible, and the Big Data Working Group created the Knowledge Hub https://bigdatahub.uvt.ro/ to focus on the development of big data analysis for labour market issues. The Knowledge Hub is a collaborative platform for mutual exchange and learning that was set up last year and it is the place where all the members of the ENRLMM can look for guidance when aiming to use big data in their labour market monitoring projects. It is an easily accessible source of information on the techniques used by other reference labour market observatories for a particular topic or challenge.

As part of the Knowledge Hub and in order to allow regular exchange and interaction, we have launched a series of Seminars so that we can deepen the Network's knowledge on how to use Big Data for labour market research and consulting by presenting practical cases and demonstrations.

So this is the first seminar of the series and today we count with the participation of two researchers from the West University of Timisoara and more specifically the Research Group on Social and Economic Complexity: Claudiu Brandas and Ciprian Panzaru. They will present their research and they will explain how they have used the opinion mining of job seekers' reviews as a new source of LMI.

Presentation by Claudiu Brandas and Ciprian Panzaru.

Exploring new sources of Labour Market Intelligence: job seekers sentiment analysis

The presentation of today is about exploring how sentiment analysis (more specifically job seekers' sentiment) could be used to understand some characteristics and specificities of the labour market.

Labour Market Analysis in the Big Data Era

Before we go into the sentiment analysis we will first make an introduction to the labour market (LM) analysis in the big data era.

We start with the advantages of using real time big data in labour market analysis.

- Real time analysis
- Large potential for analysis of labour market dynamics
- Optimal labour market policies

As for the sources that are used in the Big Data era to analyse the LM, the most well-known sources are job portals, which are very useful. In addition, we have web content (news media, news articles, e-commerce, bibliographic databases). Also, social media platforms and blogs. Another source is the online information (web usage and content as a sensor of human intent, sentiments, perceptions (e.g., for understating career expectations)). Another example is web searches databases, search engines trends and analytics such as google trends. Another valuable source is the information actively produced or submitted by users through mobile phone-based surveys and/or apps. Different operational metrics and other RTD (e.g., stock levels, school attendance, trainings, work permits, work contracts etc.). Or information about purchases (in-store and online credit cards) and financial transfers, could also be useful. This is a selection of sources but the limit is our imagination, creativity and expertise to access sources from which to obtain information. We have access to “unlimited” resources.

What can we do with this information collected from the different sources? Some examples and topics that could be analysed are:

- Labour demand and supply;
- Unemployment forecasting;
- Employment and earnings;
- Emergence of new occupations;
- New skills required on labour market;
- Skills needed by occupation;
- Post graduate tracking;
- Career expectations.

The use of big data for analysing these topics is not new and we already count with many research projects and examples conducted by different organisations and institutions. For example, regarding the use of LMI from online job advertisements (OJA), our colleagues from Milano Bicocca are obtaining information about location, occupation, educational level, salary, skills, sector, type of contract, working hours etc. from the online vacancy analysis (CEDEFOP/CRISP). This is a good source of information and probably the most well-known.

Another example is LMI from IP location. We have some studies from 2012 and 2013 to estimate and predict short- and medium labour migration flows through the Internet protocol (IP) addresses of website logins and sent e-mails. The IP addresses were used to map the geographic locations from where 43 million anonymized users sent e-mail messages within a given period. (Zagheni and Weber).

The next example is about LMI from social media: Career histories provided by LinkedIn, labour mobility using geo-located posts on Twitter or using aggregated, anonymized data on Facebook users who list their professional status on their Facebook profile. The use of these sources depends on the ability to access this data. Social media is difficult to scrape. https://economicgraph.linkedin.com/ for example, provides data about who is hiring, what jobs are available, and what skills are required in more than 180 countries.

Regarding the use of LMI from online search, recently, Italian authors have used Google search to make estimations of labour mobility or unemployment using Google Trends (analyzing queries introduced in Google Search). (Naccarato, A., Falorsi, S., Loriga, S., & Pierini, A.)

And the last example is related to LMI from mobile apps. The concept of mobile analytics involves measuring and analysing data generated by mobile applications. We can access data collected from the users, including profile information such as age and gender, location or search history can be analysed in Real-Time. (Facebook Jobs, Glasdoor, Indeed).

Sentiment analysis

In the research conducted by the West University of Timsoara, they have tried to see if online text related to LM could be used as a source of LMI using sentiment analysis. Sentiment analysis is the process of determining whether a piece of writing is positive, negative or neutral. Where does this online text come from? It could come from social media, surveys, employee interaction data, news articles and videos (sentiment analysis can be applied not only to text but also to video or audio) or from employees' reviews. In this research, they have focused on the content generated by employees' reviews, and more specifically reviews extracted from the Indeed job portal.

Before presenting the results of the research, some extra information about sentiment analysis should be said:

- is a part of "Text Analytics";
- identifies attitude (positive, negative, neutral) and emotion (joy, sadness, anger, pleasure, disgust, hate, surprise) from a message/ comments (text, voice or video);
- using AI (Artificial Intelligence) algorithms for attitude and emotion classification. The most used is NLP (Natural Language Processing) algorithms such as: Support Vector Machines, Bayesian Networks, Deep Learning.

The common tools that could be used for sentiment analysis are:

- MS Azure Machine Learning Text Analytics
- Google Natural Language API
- IBM Watson Natural Language Understanding Text Analysis
- Python with NLTK / TextBlob / Keras

In the Romanian case they have used MS Azure to analyse the job reviews of the Indeed job portal.

The research model was developed in 4 stages. The first stage, Processes, which means Data extraction, Text pre-processing and finally Sentiment detection. The second stage, Techniques, we use Web scraping for data extraction, then Translation, Tokenisation, Stemming, and Normalisation and finally Sentiment classification. Most of the content that was analysed was in Romanian language and most of the tools are designed exclusively for English language, so content had to be translated into English. The third stage, the Tools used for extracting the data from the Indeed job portal, a script was created in Python. The review was extracted, also the location and the professional position of the person reviewing as well as the rating (score they provided in a scale 1-5). For processing the text, we used MS translator, RapidMiner and MS Azure. And for detecting the sentiment we used Machine Learning-TA. Sentiment Analysis using MS Azure Cognitive Service applies sentiment labels to text, which are returned at a sentence and document level, with a score for each. The labels are “positive”, “negative”, and “neutral” and the scores range from 1 to 0. The last level is about the outputs we obtain: after extracting data we obtain the Raw data, after processing we obtain the Filtered structured data, and finally to understand our findings we visually present them.

The results of the research are presented (slide 12 of the presentation). They visually presented the overall sentiment, the sentiment by ranking, the keyword cloud, the sentiment by occupation, the sentiment by status and the sentiment by regions.

With regard to the sentiment by ranking, the users allocated a score to the job they were rating and it is worth noting that the users that highly scored a job also had more positive comments, however, there is also a high share of negative sentiment in those highly rated reviews, which is curious and we have not found an explanation because this is just an experiment. Another interesting note is that occupations that tend to be more high-skilled, the sentiment that they transmit is more positive than in the occupations that require low-level skills. Also current employees tend to be more positive in their comments than former employees. With regard to the sentiment by regions, correlations were sought to see if it could depend on GDP level at county level but nothing was found. There is room to explore why we have this distribution.

In a second step, an example on how sentiment analysis could be used not only to understand the labour market but also to see the relationship with other variables. Therefore, correlations were sought between the sentiment score and the social and cultural behaviour variables such as migration rate, electoral participation, birth rate and unemployment rate. No correlation was found except in the case of unemployment rate, where a strong correlation was found: the lower the sentiment score is, the higher the unemployment rate is.

The conclusions of the study were presented split in two:

Conclusions for human resources' analytics:

- Detect and understand employees feelings is very important (e.g., concerning new workplace policies, changes in rewards and benefits, the workplace culture, etc. improve hiring and recruiting practices, training and development needs
- Contributes to data driven decision making process (e.g., identify triggers for positive and negative sentiment and identify where improvements could be made help to build and maintain an engaged workforce and increase productivity).

Conclusions for Policy Analytics:

- Contributes to data driven decision making process (develop various indicators that serve as "early warnings" on relevant topics e.g., migration trends, unemployment etc.);
- Monitoring public policies and public investments (e.g., develop the public transport infrastructure for commuters).

Finally, it is worth noting that sentiment analysis has its limitations:

- Less procedures/methods to check data accuracy/data quality (genuine reviews? what if these reviews are ALL scripted?, posted by trolls?);
- Sentiment analysis systems trained on English data exclusively, through translation, accuracy could be lost;
- Automatic sentiment analysis of reviews using machine models are less accurate trained on review data related to labour market.

Comments by Moreno Baruffini

Moreno Baruffini *(Università della Svizzera Italiana in Switzerland)*

Comment: Unlimited resources of big data, are they really unlimited? He doesn't think so. Models and techniques are needed to extract data from these resources. He thinks that we should work on this and define a clear model.

Questions:

1. A classic question, is data representative? Are we able to link the results to other data sets?
2. Results linked to the regional differences. You tried to find some correlations but you did not find any. What about linking the results to the opportunities to apply for jobs in different locations? Commuting?

About representativeness, Ciprian Panzaru forgot to mention that 8,000 reviews were analysed in the scope of this study. Reviews were analysed only from a single portal and reviews were extracted at a given moment. So we can't say that the data is representative. However, the discussion about representativity concerns not only this study but the context of big data in general. From his point of view, big data analytics is at a different level than classic statistics, so it is not necessary to talk about representativity in this big data context.

Concerning the regional differences, the reviews were classified according to the location and using classical statistics that were available at a county level, they could not find relationships. Regarding commuters, in Romania commuting is not popular but at a European level it is very important and finding correlations in this area could help policy makers.

OPEN DISCUSSION

Novelty of the approach

- Sentiment analysis is a very interesting exercise
- A very different approach from what we are used to
- It could be useful to identify early indicators or warnings
- The PES could be very interested in such an approach
- It is a bottom-up approach, it is a way of theory building
- It is a chance to open our minds and think in a different way
- Has a great potential
- Helps build new ideas
- It links what is happening in real life
- Opens the doors to new topics- what is relevant to employees

Human resources

- Labour market policies are getting strong in the side of the employee because of the demographic change
- To attract talent, we need to create better work conditions
- It would be great to escalate this study and see if we can identify policies to improve the commitment of employers with employees
- We can use this experience as a way to progress in the area

Quality of the data

- We need to change our mindsets in relation to the representativity of data
- Representativity is not the most important aspect in big data projects
- Here, it is more critical the fake identities and the reviews posted by trolls and machines, not humans

- There are other types of Bias, for example:
  - Digital divide
  - Certain occupations have more reviews than others
  - Some people are not reviewing. There are two types of reviewers: the happy and the angry ones. There could be a bias because we do not have the middle group.
  - Many reviews came from big companies, multinational companies

Impact of translation

- Something could be lost in translation and that could be the reason for having negative sentiments in the positively ranked reviews and vice versa.
- In the translation process, the accuracy could be lost.
- Not all comments were translated from Romanian to English, there were comments in English, Italian, German, Spanish.
- The solution is to develop our own model and train our own machine to understand the content posted by users

Improving this research

- The analysis could be improved by adding Dimensions to the reviews, e.g. salary, work atmosphere, etc. So that we can see which dimensions are positively or negatively evaluated.
- In the processing of the text, we could identify the topic and then analyse the sentiment in relation with these topics.
- To do so, the algorithm must be trained to identify the domain
- It could be a specific project, to develop a specific tool for the labour market context
- Another project to validate results

Fostering partnerships in research

- We could use this type of research to foster fruitful partnerships between quantitative and qualitative data
- Gives a snapshot
- Builds an understanding to see if you can/ should research further
- How can an analysis like this help build stronger partnerships in research?
- It sparks ideas

Using location

- Using location of users is an exciting area
- Could be used for example, to detect digital nomads, moving the work elsewhere
- Analysis and correlation with motivation

- Use this sort of analysis to reach freelancers, who usually escape the classical statistics
- Locate them and see where they are and make analysis about what we find out: if they live in clusters, if they are enriching or impoverishing the community…

Next steps

- We need more sessions of this kind
- Hands-on sessions so that we can learn how to carry out this kind of projects
- Further research like this one
- The project could be extended at a network level and develop indicators at a European level or a ENRLMM index aggregating information based on sentiment analysis of online text
- Protecting users and individuals in the wider digital areas, important for shaping the proper protocols in research

Closing statement

This was an exercise to spark new ideas within the network

Bilbao, April 2022

References
Speaker: CIPRIAN PANZARU
Universitatea de Vest din Timișoara, Romania
Email: ciprian.panzaru@e-uvt.ro
EUROPEAN NETWORK ENRLMM
Email: jenny@jennykipper.de
Email: c.larsen@em.uni-frankfurt.de
Website: www.regionallabourmarketmonitoring.net
BIG DATA WORKING GROUP
Eugenia Atin
E-mail: e.atin@prospektiker.es
Website: https://bigdatahub.uvt.ro/