Statistical scraping – related initiatives in Eurostat

Statistical Scraping Interest Group meeting, 16 September 2025

European Commission

LLM-based classification of OJAs

Eurostat's OJA data collection

In the framework of the Web Intelligence Hub, Eurostat has for several years been collecting Online Job Advertisements (OJA) from job portals to provide additional insight on labour market dynamics. This has resulted in the publication of experimental statistics:

- Online job advertisement rate
- Labour market demand for ICT specialists in OJAs

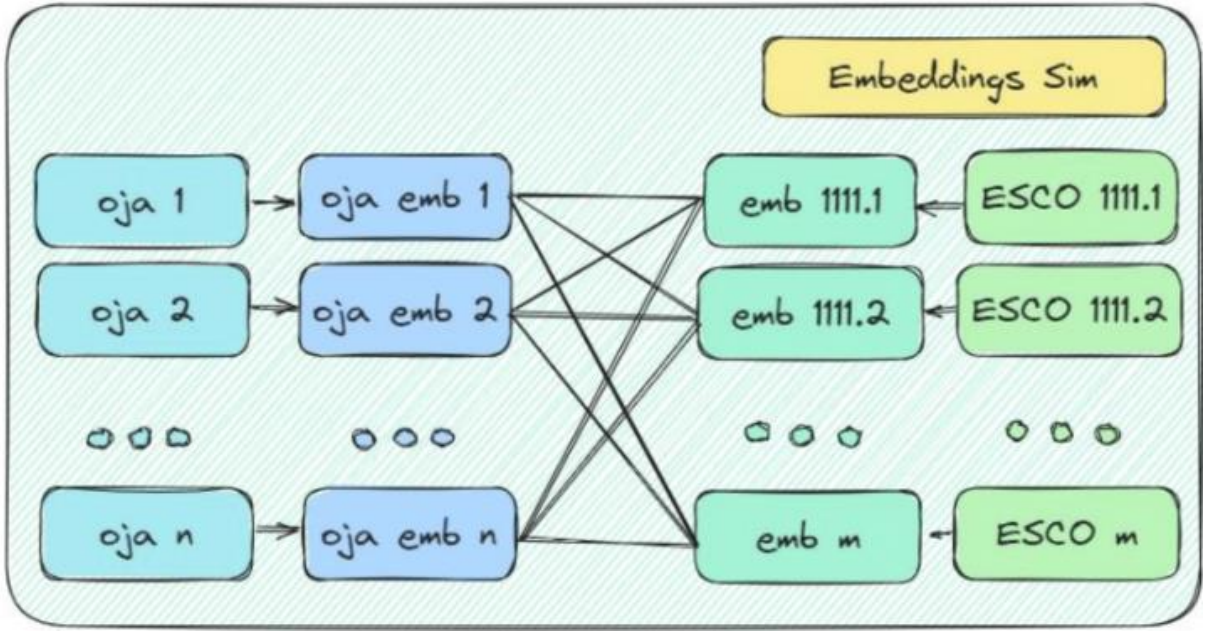
The approach does not fit the "Statistical scraping" concept – we cast a wide net rather adopting a targeted approach.

One of the challenges in this approach is to correctly classify job advertisements according to the available classifications for occupations (ISCO and ESCO).

Use of LLM's for classifying OJAs

The classification of OJAs has relied on "traditional" NLP methods (e.g. matching job titles / descriptions to taxonomy items using available ontologies or ML models).

Eurostat has conducted a study on the usage of LLM methods for this purpose instead, and in particular on the usage of embedding models.



ESCO occupation descriptions are vectorised using available embedding models.

OJA job descriptions also vectorised, then best match found.

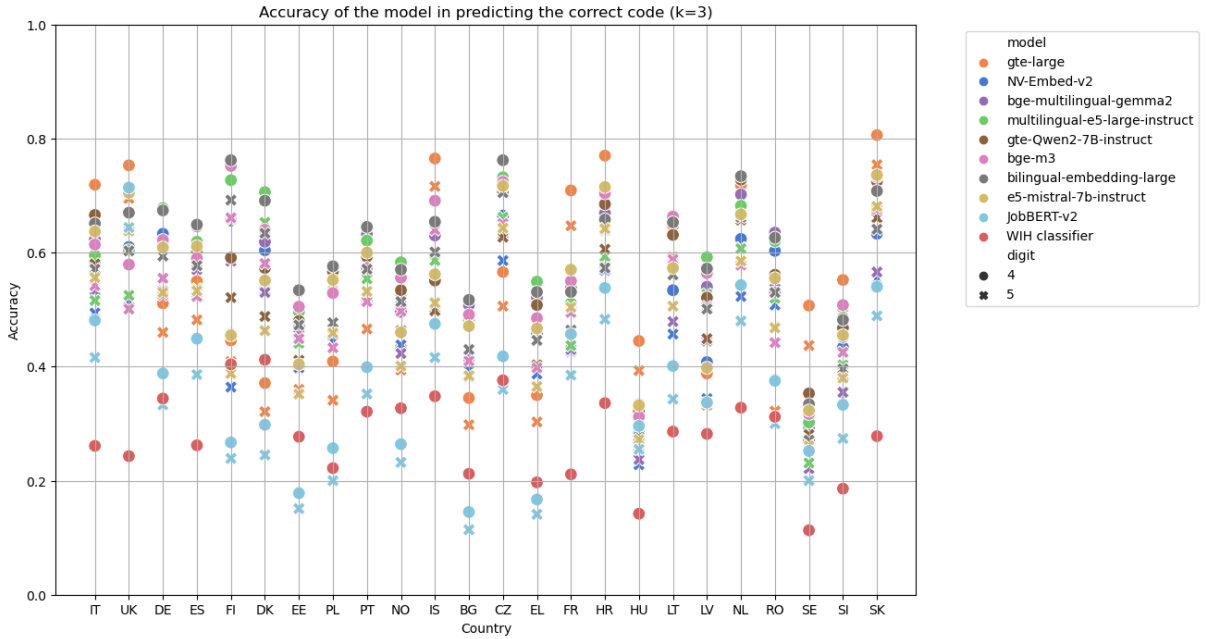
Structure of the study

- Select top-tier LLMs based on the “Massive Text Embedding Benchmark” (MTEB)
- Select a sample of OJAs using Farthest Point Sampling (FPS). This ensures a diverse selection of job classifications, preventing the overrepresentation of certain categories.
- Sample OJAs are classified to the correct classification item by labour market experts.
- Classify the OJAs using the selected embedding models
- Measure the accuracy of the classification procedure by applying a specific similarity metric that gives a higher weight to larger classification errors.

Results and next steps

Different embedding models were tested. The accuracy of the classification varied among them, but they in general outperformed the traditional NLP methodology used so far.

Eurostat is now working on verifying the robustness and sustainability of the approach with the aim of implementing the new embedding-based classification in its OJA production process. The outcomes of the study will be made publicly available.



MNE data from Wikipedia

Objectives

Extract information about “top-tier” multi-national enterprises (MNEs) from Wikipedia in order to validate / supplement information available in business registers:

- Number of employees
- Net income
- Revenue
- Total assets
- Website
- Country where headquarters located

“Top-tier” MNEs: list of around 1100-1200 important MNEs defined by the Statistical Business Registers Working Group

Wikipedia as a data source

✓ Key information about enterprises available in a somewhat standardised format (via the available info box) – reduced number of scrapers needed!

✓ Scraper-friendly environment from the technical and legal perspective - scraping is not blocked, information is not covered by copyright (CC0 license)

~ No certainty about quality or “freshness” of figures - though the source and reference year for key enterprise figures is often provided!

~ Does not cover SMEs – but provides very good coverage for large “top-tier” MNEs

Company type	Private company
Industry	Aviation, tourism
Founded	2010[1]
Headquarters	Dublin, Ireland
Key people	Gediminas Žiemelis (chairman of the board) Jonas Janukenas (CEO and member of the board)
Revenue	€2.263 billion (2023) [2]
Operating income	€160 million (2023) [2]
Number of employees	11 000+ (2023)
Rating	Fitch Ratings – BB S&P Global Ratings – BB-
Website	www.aviasg.com

Finding the Wikipedia pages

The following procedure was used to find the Wikipedia pages corresponding to top-tier MNEs:

- Landscaping of Wikipedia categories likely to contain pages related to top-tier MNEs
- Extraction of the titles of all pages available in these categories
- Matching page titles to available names of top-tier enterprises via string distance algorithms

This procedure allowed finding the relevant Wikipedia pages for around 50% of top-tier MNEs. A manual procedure was employed for the rest.

Current status and future outlook

Wikipedia figures for top-tier enterprises are now extracted quarterly and shared with members of the Statistical Business Registers Working Group.

Eurostat is also considering attempting to extract information from financial reports made available by top-tier MNEs on their websites. Compared to Wikipedia:

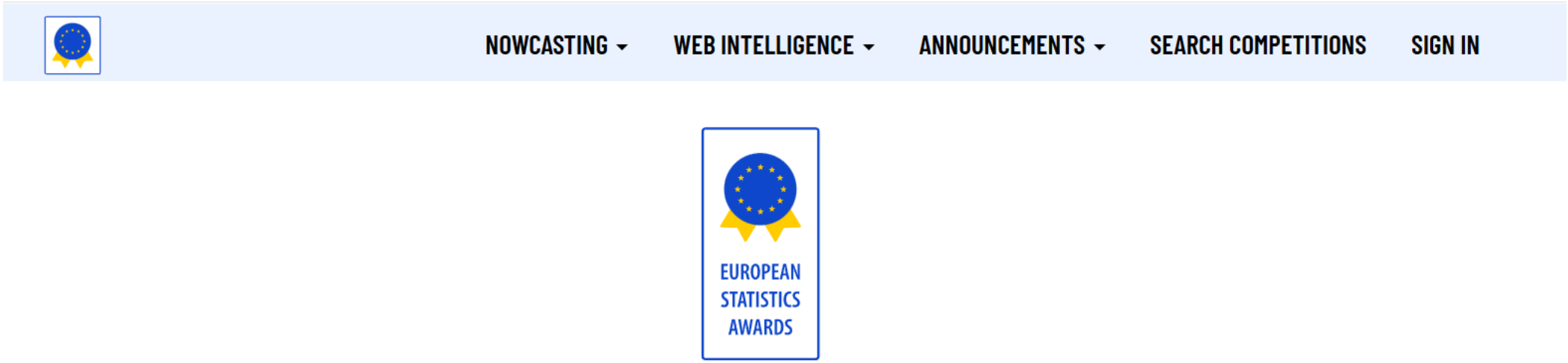
- Likely higher data quality
- More sophisticated algorithms are needed to find the reports
- Higher degree of customisation for scraping and extraction tools, as financial reports for different enterprises have different formats / structure.

Other initiatives

European Statistics Awards

The European Statistics Awards are a set of competitions launched by Eurostat to "crowdsource" the development of new solutions to improve official statistics.

The European Statistics Awards is composed of competitions on two strands of work: Nowcasting and Web Intelligence. For each competition, monetary prizes are given to the top 3 teams in three categories: Accuracy, Reusability and Innovativity.



European Statistics Awards

In the Web Intelligence strand, competitions have been launched on the following topics.

Closed competitions:

- Deduplication challenge: identifying potential duplicates of job postings published on the web.
- Classification challenge: developing approaches that learn how to assign a class label (from a known taxonomy) to job advertisements from a given dataset.

Competitions under evaluation:

- Discovery challenge: develop approaches that automatically identify sources of annual financial data on the internet for MNE groups.
- Extraction challenge: develop approaches that automatically extract important annual financial data of MNE Groups.

Eurostat will share summaries of the methodologies / approaches proposed in these competitions.

Clarification of legal framework for scraping

In 2017, as part of the Big Data ESSnet, a review of the legal framework around web scraping had been conducted.

Since then, several new relevant legal acts have appeared at European level:

- Directive (EU) 2019/790 on Copyright in the Digital Single Market contains important provisions regarding exceptions to intellectual property rights in case of text and data mining.
- The amended regulation 223/2009 introduces new legal provisions which may facilitate access to privately held data

Eurostat is therefore working on providing an updated overview of the European legal framework for web scraping.

In case of European **directives**, national transposition details may also be very important

Updated landscaping for OJA

In 2025, Eurostat will launch an update of its landscaping exercise for job portals.

The methodology adopted during the last landscaping exercise in 2021 will be adapted. Comments and recommendations provided by the WIN ESSnet will be taken into account.

Fill in copyright information.

Thank you