Classification Approach Description

Methodology *(Replicability; Scalability; Interpretability)*

Please provide a detailed description of the methodology used for identifying the classification of the online job advertisement. The description should contain (1) the data processing steps, (2) the methods and models used, (3) references to the scientific papers/sources that present the methods and models used, and (4) the time it took to process the data set and classify the job advertisements.

Bear in mind that the workflow will be also evaluated based on the criteria for the Reusability and Innovativity Awards.

This section will be evaluated for:

> *(1) the Replicability criterion: likeliness that the described approach can successfully reproduce the solution submitted by the team for the Accuracy award*
>
> *(2) the Scalability criterion: amount of modification required for the approach to apply to similar datasets on a potentially larger scale*
>
> *(3) the Interpretability criterion: the extent to which a human could understand and articulate the relationship between the approach's predictors and its outcome; how well the logical reasoning behind the model which is making the prediction is developed (whether it is mathematically and/or technically sound*

Overview

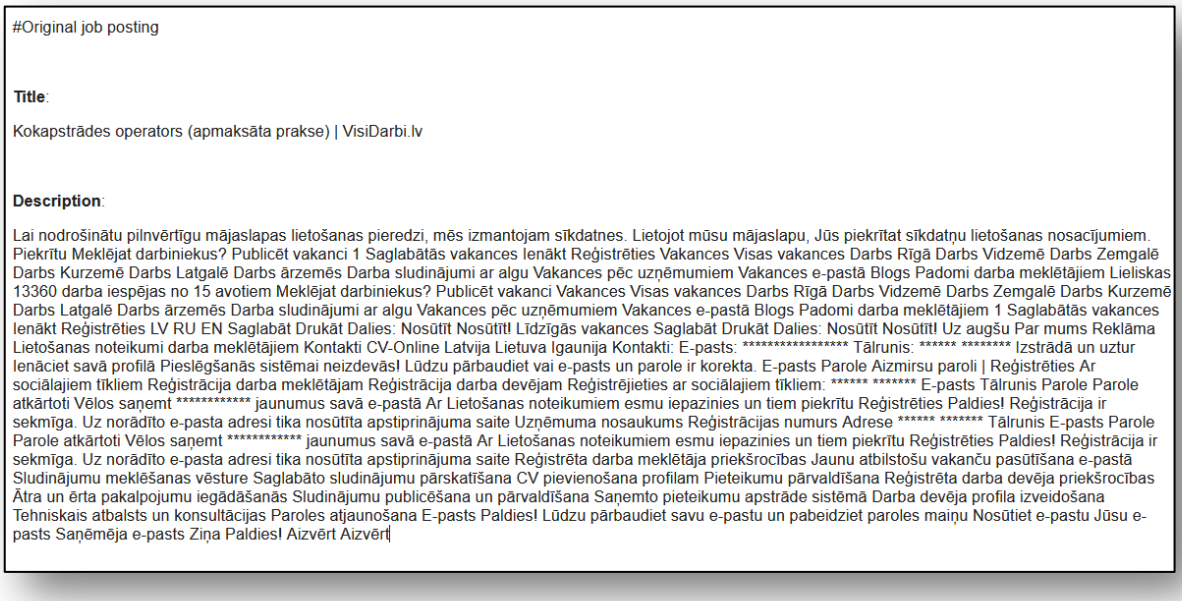
The classification challenge involves identifying the ISCO category of each job advertisement. To achieve this, we downloaded the ESCO dataset from the Eurostat website and merged several tables to create a comprehensive dataset that includes occupations, occupation descriptions, required skills, and more.

Once the data was collected, we built a retrieval system using **Weaviate**, which allowed us to perform an initial filter on the occupations and categories to obtain the most relevant ones for the job advertisements. The retrieval phase employed a hybrid search that combined both semantic similarity and keyword search. Given that the job advertisements contained many irrelevant words that hindered classification, we needed to clean and summarize them. We utilized **gemma-2-9b-it-SimPO** to extract key features such as job title, skills, industry, description, and required experience from the advertisements. Additionally, we asked the LLM to translate the ads into English to facilitate an English-only retrieval.

To classify the job ads, we implemented **Retrieval-Augmented Generation (RAG)**. Our RAG implementation used **all-mpnet-base-v2** to vectorize occupations related to the job features and bge-reranker-large to rerank the retrieved occupations. We also checked for matches between the job title and titles in the ESCO dataset; if a match was found, the corresponding occupations were added to the retrieved results. We then passed the extracted features of the

job to **gemma-2-9b-it-SimPO** (in alternative **gpt-4o-mini*** through an API could also be used) to assign a category.

***Notice:** The features sent to the API would includ a cleaned and schematic version of the job title, a cleaned summary of the job description, seniority level, and industry. All features would be stripped of identifiable information that could reconstruct the original job ads (e.g., location, company name, salary, date, etc.). Furthermore, the data sent through the API would be securely transmitted, not used for further model training, and would be deleted after 30 days, as stated on the LLM provider's website.

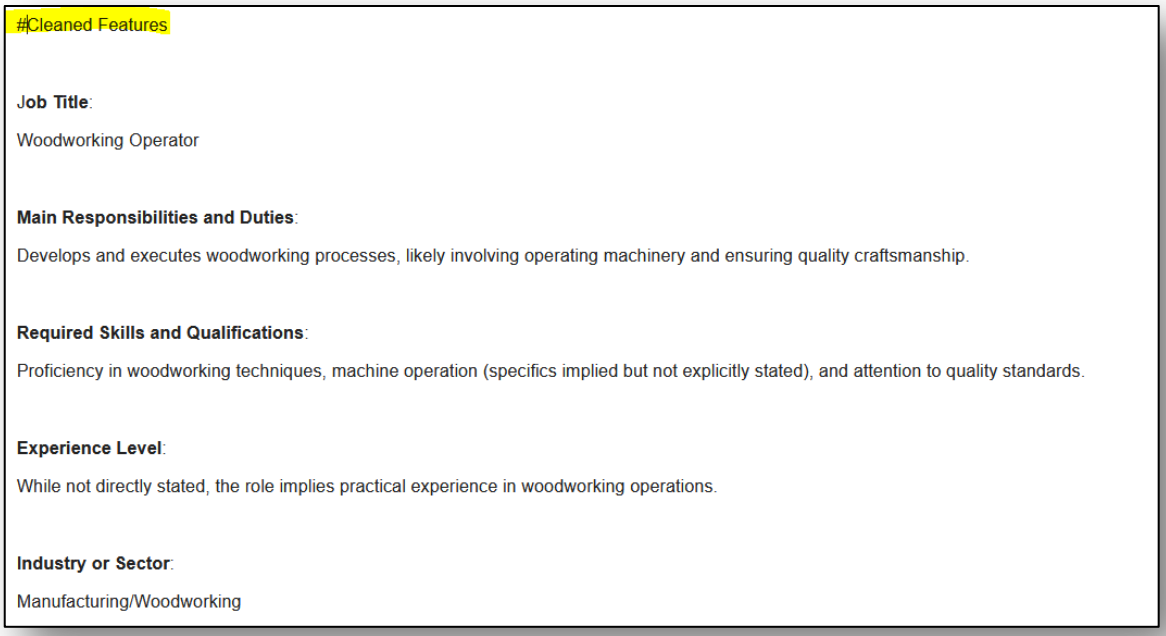

#Original job posting

Title:

Kokapstrādes operators (apmaksāta prakse) | VisiDarbi.lv

Description:

Lai nodrošinātu pilnvērtīgu mājaslapas lietošanas pieredzi, mēs izmantojam sīkdatnes. Lietojot mūsu mājaslapu, Jūs piekrītat sīkdatņu lietošanas nosacījumiem. Piekrītu Meklējat darbiniekus? Publicēt vakanci 1 Saglabātās vakances Ienākt Reģistrēties Vakances Visas vakances Darbs Rīgā Darbs Vidzemē Darbs Zemgalē Darbs Kurzemē Darbs Latgalē Darbs ārzemēs Darba sludinājumi ar algu Vakances pēc uzņēmumiem Vakances e-pastā Blogs Padomi darba meklētājiem Lieliskas 13360 darba iespējas no 15 avotiem Meklējat darbiniekus? Publicēt vakanci Vakances Visas vakances Darbs Rīgā Darbs Vidzemē Darbs Zemgalē Darbs Kurzemē Darbs Latgalē Darbs ārzemēs Darba sludinājumi ar algu Vakances pēc uzņēmumiem Vakances e-pastā Blogs Padomi darba meklētājiem 1 Saglabātās vakances Ienākt Reģistrēties LV RU EN Saglabāt Drukāt Dalies: Nosūtīt Nosūtīt! Līdzīgās vakances Saglabāt Drukāt Dalies: Nosūtīt Nosūtīt! Uz augšu Par mums Reklāma Lietošanas noteikumi darba meklētājiem Kontakti CV-Online Latvija Lietuva Igaunija Kontakti: E-pasts: ********** Tālrunis: ****** ******** Izstrādā un uztur Ienāciet savā profilā Pieslēgšanās sistēmai neizdevās! Lūdzu pārbaudiet vai e-pasts un parole ir korekta. E-pasts Parole Aizmirsu paroli | Reģistrēties Ar sociālajiem tīkliem Reģistrācija darba meklētājam Reģistrācija darba devējam Reģistrējieties ar sociālajiem tīkliem: ****** ******* E-pasts Tālrunis Parole Parole atkārtoti Vēlos saņemt ********** jaunumus savā e-pastā Ar Lietošanas noteikumiem esmu iepazinies un tiem piekrītu Reģistrēties Paldies! Reģistrācija ir sekmīga. Uz norādīto e-pasta adresi tika nosūtīta apstiprinājuma saite Uzņēmuma nosaukums Reģistrācijas numurs Adrese ****** ******* Tālrunis E-pasts Parole Parole atkārtoti Vēlos saņemt ********** jaunumus savā e-pastā Ar Lietošanas noteikumiem esmu iepazinies un tiem piekrītu Reģistrēties Paldies! Reģistrācija ir sekmīga. Uz norādīto e-pasta adresi tika nosūtīta apstiprinājuma saite Reģistrēta darba meklētāja priekšrocības Jaunu atbilstošu vakanču pasūtīšana e-pastā Sludinājumu meklēšanas vēsture Saglabāto sludinājumu pārskatīšana CV pievienošana profilam Pieteikumu pārvaldīšana Reģistrēta darba devēja priekšrocības Ātra un ērta pakalpojumu iegādāšanās Sludinājumu publicēšana un pārvaldīšana Saņemto pieteikumu apstrāde sistēmā Darba devēja profila izveidošana Tehniskais atbalsts un konsultācijas Paroles atjaunošana E-pasts Paldies! Lūdzu pārbaudiet savu e-pastu un pabeidziet paroles maiņu Nosūtiet e-pastu Jūsu e-pasts Saņēmēja e-pasts Ziņa Paldies! Aizvērt Aizvērt



#Cleaned Features

Job Title:

Woodworking Operator

Main Responsibilities and Duties:

Develops and executes woodworking processes, likely involving operating machinery and ensuring quality craftsmanship.

Required Skills and Qualifications:

Proficiency in woodworking techniques, machine operation (specifics implied but not explicitly stated), and attention to quality standards.

Experience Level:

While not directly stated, the role implies practical experience in woodworking operations.

Industry or Sector:

Manufacturing/Woodworking

Steps and procedures

1) **Build the dataset:**
   After downloading the ESCO dataset, we merged the tables to obtain the following for each occupation:
   - Title
   - Alternative titles
   - Description
   - Skills
   - Category title and description

   Execution time: **less than 1 minute**

2) **Translate, Clean and Extract Job Features:**
   We constructed a prompt to ask **Gemma-2-9b-it-SimPO** to translate the job advertisements into English and to clean them by removing irrelevant content. The extracted features included:
   - Job title
   - Job Description
   - Skills Required
   - Experience
   - Industry

Execution time: **24 hours on NVIDIA H100 80GB HBM3 (less if quantized model is used)**

Original job posting

Title:
Panel & Paint Technician

Description:
Panel & Paint Technician required in Colchester, Essex Hours: Monday to Friday – 45 hours 7.00am-4.30pm Basic Salary: Up to £35,000 basic p.a. (depending on experience and qualifications held). Bonus: Time saved bonus up to £600 per month. Overtime: Saturdays optional. Company: Well established Accident Repair Centre. A fantastic opportunity has arisen within our Motor Trade clients business for a qualified and experienced Panel & Paint Technician to join their busy Motor Trade Bodyshop. You will be able to work on your own initiative and be capable of producing a high standard of workmanship. What our clients are looking for in our panel and paint technician. Completing small jobs such as bumper and wing changes. Able to complete Light MET/Panel & Paint work. Good at filler work. You will ideally be City & Guilds & ATA qualified. This is a great opportunity to join this well established and progressive automotive group. If your time has come, you have drive, ability and the desire to be an important and key part of this well rewarded team. Motor trade experience is essential to the role. If this Panel and Paint Technician Job interests you and you would like to know more about it or other Automotive Jobs in Colchester please contact ****** ***** at Perfect Placement UK Ltd. At Perfect Placement, we specialise in Motor Trade Jobs, so if you are looking to improve your Automotive Career please get in touch with us today. We are the leading UK Motor Trade Recruiters with over 1,400 Automotive Vacancies across the whole of the UK!!

Extracted Features:
Job Title:
Panel & Paint Technician

Description:
A qualified and experienced technician is sought to perform light automotive body repair and paintwork, including bumper and wing replacements, filler work, and MET/panel repairs. City & Guilds and ATA qualifications are desirable, along with proven motor trade experience. The role emphasizes high-quality workmanship and independent work execution.

Job Title:
Panel & Paint Technician

Main Responsibilities and Duties:
Completing small repairs (bumper, wing changes), light MET/Panel & Paint work, filler work, producing high-quality workmanship independently.

Required Skills and Qualifications:
City & Guilds & ATA qualifications, proven Motor Trade experience.

Experience Level:
Experienced (explicitly mentioned)

Industry or Sector:
Automotive/Motor Trade

Keyword feature
Panel & Paint Technician, Light MET/Panel Repair, Filler Work, City & Guilds, ATA Qualification, Motor Trade Experience

3) **Build Weaviate Collection:**
We utilized **Weaviate** to index and vectorize each occupation. The vectorization was performed using the **all-mpnet-base-v2** model. The search engine employed a hybrid approach, combining semantic search and keyword search, with particular emphasis on the importance of the job title.
Execution time: **less than 5 minutes**

4) **RAG+rerank:**
For each job, we employed the retrieval system to find the most relevant occupations. The search engine (Weaviate) includes a parameter to adjust the balance between the two search methodologies (i.e., only keyword or only vector similarity). The parameter alpha ranges from 0 (keyword search only) to 1 (vector similarity only). For each query, we retrieved the 10 most relevant occupations at alpha values of 0, 0.5, and 1.

Next, we used **bge-reranker-large** to rerank the retrieved occupations and appended the top 10 results with those obtained through perfect title matching, if any.

We then grouped the occupations according to their corresponding 4-digit ISCO categories, creating short text descriptions formatted as follows:

```
ISCO Category: [label of ISCO category]
ISCO category code: [4-digit unique ISCO code]
ISCO Category Description: [description of the ISCO category]

Occupations in this category that might be related to the job ad:

Occupation Title: [title of the 1st occupation retrieved]
Occupation Alternative Titles: [alternative possible title of the 1st occupation
retrieved]
Occupation Description: [description of the 1st occupation retrieved]

Occupation Title: [title of the 2st occupation retrieved]
Occupation Alternative Titles:[alternative possible title of the 2nd occupation
retrieved]
Occupation Description: [description of the 2nd occupation retrieved]

....
```

Finally, we concatenated all these descriptions into a prompt along with the job posting to be classified. We then asked the LLM to assign each job posting to a unique ISCO category code.

Execution time: **10 hours on NVIDIA H100 80GB HBM3**

Summary

This methodology emphasizes the following criteria:

Replicability: The approach can be easily reproduced by following the detailed steps outlined above, utilizing the specified models and techniques. By providing a clear breakdown of the process and execution times, other researchers can replicate the results with similar datasets.

Scalability: The methodology is designed to be scalable, allowing for modifications to apply to larger datasets. By using open-source models and a flexible retrieval system, the process can accommodate additional job advertisements without significant changes to the underlying framework.

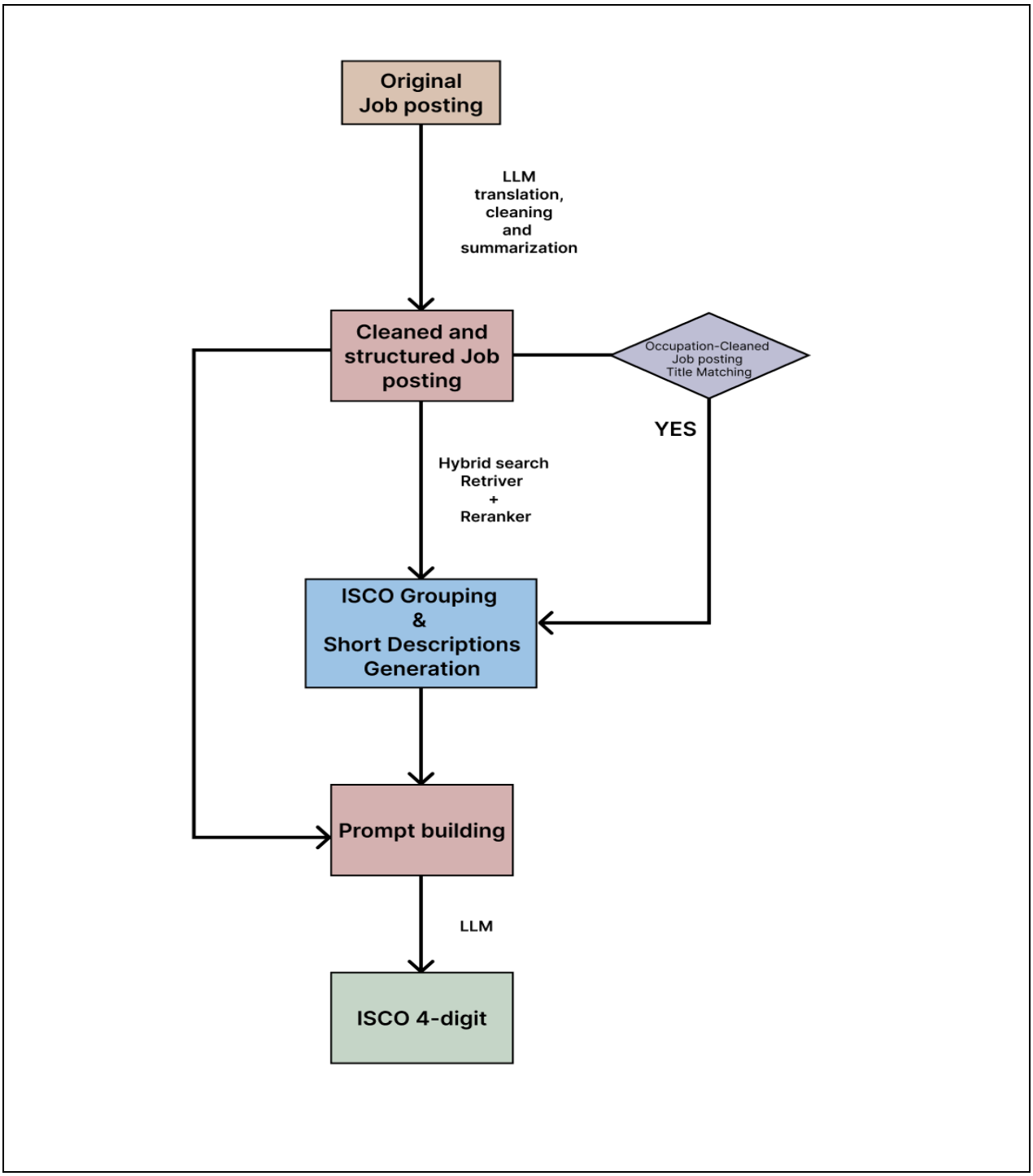
Interpretability: The approach is structured to ensure that the relationships between the extracted features and the assigned ISCO categories are transparent. Each step is logically sequenced, making it easier for humans to understand how job features correlate with their respective classifications. The use of hybrid search and the clarity of the RAG implementation contribute to a solid technical foundation that enhances interpretability.

Architecture

Please provide a description of the architecture of your approach. A diagram of the architecture is considered of additional value. Indicate what modifications would be required to apply the approach to similar datasets on a larger scale.

This section will be evaluated for:

(1) *the Architecture criterion: evaluated based on its modules, their cohesion and their configurability; an architecture which is modular and includes clear connections between modules or components receives a higher score*


Original Job posting
LLM translation, cleaning and summarization
Cleaned and structured Job posting
Occupation-Cleaned Job posting Title Matching
YES
Hybrid search Retriver + Reranker
ISCO Grouping & Short Descriptions Generation
Prompt building
LLM
ISCO 4-digit

Hardware Specifications *(Replicability; Scalability; Interpretability)*

Please describe the hardware specifications of the machines that were used to run the methodology.

This section will be evaluated for:

(1) the Replicability criterion

(2) the Scalability criterion

(3) the Interpretability criterion

Machine 1

CPUs	Physical cores: 112 Logical cores: 224
GPUs	NVIDIA H100 80GB HBM3 with 81041.75 MB of memory
TPUs	NONE
Disk space	Total memory: 2015.56 GB

Machine 2

CPUs	Physical cores: 8 Logical cores: 8
GPUs	GPU MPS: Apple GPU (MPS) with 8GB of memory
TPUs	NONE
Disk space	Total memory: 8.00 GB

Libraries *(Maintainability)*

Please provide the libraries used for approach, if any, as well as the links to these libraries, if available.

This section will be evaluated for:

(1) the Maintainability and openness criterion: use of libraries which are regularly maintained will yield higher scores. (Examples include pytorch, tensorflow, scikit-learn, pandas, numpy, etc.) The use of libraries which are openly available will yield higher scores.

- numpy
- pandas
- os
- re
- json
- torch
- transformers
- FlagEmbedding

- sentence_transformers
- OpenAI (optional)
- weaviate

Open license *(Maintainability)*

Please provide the open license of the provided code, if any.

This section will be evaluated for:

> *(1) the Maintainability and openness criterion: whether the approach is open and under an open license*

Similarities/differences to State-of-the-Art techniques *(Originality)*

Please provide a list of similarities and differences between the used methodology and to the state-of-the-art techniques.

This section will be evaluated for:

> *(1) the Originality of the approach criterion: compare the approach used to the state-of-the-art; the extent to which the submission represents an improvement over these pre-existing approaches*

- **LLM:**
  Large Language Models (LLMs) are currently the state-of-the-art for Natural Language Processing (NLP). We utilized an open-source model with only 9 billion parameters to extract features from text, employing prompts based on best practices in prompt engineering.
- **RAG:**
  LLMs are known to be prone to hallucinations. When we asked the LLM to assign categories to job ads, the model generated responses that demonstrated some knowledge of ESCO, but the assigned categories were mostly incorrect or non-existent. By using Retrieval-Augmented Generation (RAG), we enabled the LLM to select from existing categories relevant to the job ads. Specifically, we implemented a hybrid search that allowed us to weigh certain features more heavily than others (e.g., Job Title) while performing keyword searches.

Contribution to scientific field *(Future orientation)*

Please describe how your submission contributed to the scientific field, what impact it could have and what could potentially be future work to improve the solution.

This section will be evaluated for:

> *(1) the Future orientation and impact criterion: the potential effect of the approach used will be evaluated; this includes the scale of impact it has on the problem of the classification of job advertisements; the impact will be evaluated based on potential efficiency improvements and cost reductions.*

Our approach has the potential to significantly impact the classification of job advertisements. It is highly scalable and offers substantial room for improvement.

We developed the system in an unsupervised manner due to the absence of predefined categories. However, if we had access to real categories, we could fine-tune the large language model (LLM) to enhance its classification capabilities. While the LLM we used is designed for various tasks such as summarization and translation, it is not specialized in job classification according to the European Skills, Competences, Qualifications and Occupations (ESCO) framework. We anticipate that fine-tuning the model will lead to improved performance in this specific task.

Furthermore, we expect enhancements in the Retrieval system when utilizing actual categories to adjust prompt and parameters to maximize key metrics such as Hit Rate.

The optimized system could assist Eurostat in automating job classification, resulting in significant cost and time savings. An intriguing approach could involve human-machine collaboration, where the model learns from human input to classify simpler job ads. For more complex cases, the model could retrieve the most relevant categories, thereby streamlining the process and improving the quality of human-labeled data.

Lessons Learned *(Future orientation)*

Please state any lessons learned during the competition.

This section will be evaluated for:

> *(1) the Future orientation and impact criterion: what were the lessons learnt during the competition, and what could potentially be future work to improve the solution.*

The lessons we learned are:

- **Hybrid Search Seems to Perform Better**:
  By using hybrid search, we can give more weight to job titles and retrieve better matching occupations.

- **Some Job Ads were not Actual Jobs:**
  Many job ads were actually webinars or other types of content.

- **The Job Ads Contained Many Non-Relevant Facts:**
The job ads included information related to the job application website, such as cookie management, privacy terms, application steps, and more. Cleaning this text improved the quality of the retrieved occupations.

- **English-Only Retrieval on Translated Ads Outperforms Multilingual Retrieval on Original Ads:**
We compared the original ads using a multilingual approach with translated ads using an English-only approach. Our best results were achieved with the English-only method.

- **Classify job ad is hard! Even for humans:**
We attempted to assign categories to job ads, but it was challenging because many ads matched multiple categories. In these cases, the assigned category was primarily based on personal opinion and interpretation.

- **Open Source LLMs Is all you need:**
We evaluated the performances of various small (<=9 billion parameters) open source LLMs. The models were all quite good, the one that showed the best performances was gemma-2-9b-it-SimPO.

Short description of the Team – area of expertise

Please provide a description of the team, your area of expertise and contact information.

Andrea Alessandrelli, Msc in Physics, Phd Student
Pasquale Maritato, Msc in Economics, Data Scientist

Contacts:

- a.alessandrelli@studenti.unipi.it
  tel +39 3272268760
- p.maritato@studenti.unipi.it
  pasquale.maritato@outlook.com
  tel +39 3458032231