From Data to Insight: Transforming Online Job Postings into Labor-Market Intelligence

Giannis Tzimas 1,*, Nikos Zotos 2,*, Evangelos Mourelatos 3, Konstantinos C. Giotopoulos 2, and Panagiotis Zervas 1

1 Data and Media Laboratory, Department of Electrical and Computer Engineering, University of Peloponnese, 22131 Tripoli, Greece; p.zervas@uop.gr
2 Department of Management Science and Technology, University of Patras, 26334 Patras, Greece; nzotos@upatras.gr (N.Z.); kgiotop@upatras.gr (K.C.G.)
3 Department of Economics, Accounting and Finance, Oulu Business School, University of Oulu, FI-90014 Oulu, Finland; evangelos.mourelatos@oulu.fi
* Correspondence: tzimas@uop.gr

Abstract: In the continuously changing labor market, understanding the dynamics of online job postings is crucial for economic and workforce development. With the increasing reliance on Online Job Portals, analyzing online job postings has become an essential tool for capturing real-time labor-market trends. This paper presents a comprehensive methodology for processing online job postings to generate labor-market intelligence. The proposed methodology encompasses data source selection, data extraction, cleansing, normalization, and deduplication procedures. The final step involves information extraction based on employer industry, occupation, workplace, skills, and required experience. We address the key challenges that emerge at each step and discuss how they can be resolved. Our methodology is applied to two use cases: the first focuses on the analysis of the Greek labor market in the tourism industry during the COVID-19 pandemic, revealing shifts in job demands, skill requirements, and employment types. In the second use case, a data-driven ontology is employed to extract skills from job postings using machine learning. The findings highlight that the proposed methodology, utilizing NLP and machine-learning techniques instead of LLMs, can be applied to different labor market-analysis use cases and offer valuable insights for businesses, job seekers, and policymakers.

Keywords: labor-market analysis; online job postings; labor information extraction; NLP; machine learning; data-driven ontology

1. Introduction

The labor market has undergone a significant transformation in recent years, primarily due to the widespread use of online job postings by employers. Initially, these postings were predominantly for highly skilled personnel. However, contemporary online job portals (OJPs) now feature job offers encompassing nearly all occupation categories, skill levels, and industry sectors, representing a great data source for LM specialists to analyze the labor market [1]. Even the recruiting human resources companies utilize OJPs for job postings to attract qualified applications [2].

This shift has generated a vast amount of data, offering valuable insights into labor-market dynamics. Concurrently, the COVID-19 pandemic has significantly impacted the global economy, including the labor market. Consequently, online job postings have become increasingly crucial for businesses, job seekers, and policymakers. An analysis of online job postings data can identify trends in the labor market, including the availability of specific skills, wage trends, regional disparities, and diversity and inclusion issues.

Although these data provide great potential in labor-market analyses, processing online job postings may cause multiple challenges. These challenges include handling the vast
volume of data; dealing with inconsistencies and missing information in job postings; and accurately extracting relevant information such as job titles, required skills, and employer details. Our research addresses these challenges by proposing a comprehensive methodology for processing online job postings and transforming them into actionable labor-market intelligence. This methodology includes systematic data source selection, data extraction, cleansing, normalization, and deduplication procedures, followed by advanced Natural Language Processing (NLP) and machine-learning techniques for information extraction.

Despite the recent domination of Large Language Models (LLMs) in data analysis and information extraction, classic NLP tasks are not obsolete, but their role and usage have evolved significantly. For a labor-market analysis, classic NLP techniques may be preferable due to their interpretability and precision when dealing with domain-specific terminology and structured data. These methods can provide transparent insights into job market trends, skill demands, and industry-specific language, which is crucial for making informed policy decisions and workforce planning. Naveed H. et al. (2023) [3] provide a detailed examination of the capabilities and limitations of LLMs. While LLMs excel in numerous areas, particularly in handling large datasets and performing complex language tasks, classic NLP tasks remain highly relevant. The paper discusses scenarios where traditional NLP techniques are preferable, such as dealing with specific, well-defined problems, operating under limited computational resources, requiring interpretable and explainable models, and achieving high precision in specialized domains. This balanced perspective highlights the continued importance of classic NLP approaches alongside the advancements brought by LLMs. Singh C. et al. (2023) [4] introduced a framework for leveraging the knowledge learned by LLMs to build efficient and interpretable models. In their work, they discussed two significant issues of the LLMs performance. First, like most deep neural networks, LLMs are difficult to interpret, often being labeled as black boxes, thus hindering their application in specific domains; and second, the massive size of black-box LLMs results in high energy costs, making them expensive and challenging to deploy, especially in low-compute environments.

In our current work, we applied our methodology to two specific use cases. The first examines the Greek labor market in the tourism industry during the COVID-19 pandemic, revealing how the pandemic affected occupations, required skills, and employment types. Our analysis showed a 7.5% increase in part-time and short-term contract positions; a rise in demand for high-skilled blue-collar jobs; and evolving skill requirements, emphasizing healthcare, information management, and food service administration skills.

The second use case employs a data-driven ontology to extract skills from job postings, utilizing machine-learning techniques to develop a skills taxonomy without expert input. This approach provided a detailed understanding of the specific skills in demand, offering a real-time snapshot of labor-market trends and informing workforce development strategies.

2. Literature Review

Several studies have analyzed online job postings’ data to gain insights into the labor market’s dynamics. The European Centre for the Development of Vocational Training, commonly referred to as CEDEFOP, is an agency of the European Union that focuses on supporting vocational education and training (VET) systems through high-quality labor-market intelligence. In this field, CEDEFOP [5] conducted great research on the field of labor-market analysis, focusing on improving vocational education and training (VET) systems depending on high-quality and timely insights into labor-market dynamics and new skill demands. CEDEFOP analyzed how changing economic and social megatrends, such as workforce ageing and the digital and green transitions, generate new skill demands and skill mismatches and reshape the future of work in EU workplaces.

Specifically, in “Online job vacancies and skills analysis: A CEDEFOP pan-European approach (2019)” [6], CEDEFOP extended its well-established battery of instruments to produce skills and labor-market intelligence. It presented key features and gave a concise overview of CEDEFOP’s work, including the methodology and analytical approach, to
help understand the results. It also pointed to contextual issues and limitations that must be kept in mind when using the data. Focusing on the EU labor market, CEDEFOP in “The online job vacancy market in the EU: driving forces and emerging trends (2019)” [7] introduced the project “Real-time labour market information on skill requirements: setting up the EU system for online vacancy analysis”, in which the infrastructure was created to gather information from the most important online job-vacancy portals in real-time across the EU. When meaningfully analyzed and turned into intelligence, this information can complement skills intelligence developed using traditional methods. In addition, CEDEFOP addressed the challenges needed to be considered for collecting and analyzing online job postings and concluded that there must be a close monitoring of the online job postings and careful consideration of how they reflect the labor market-map development and emerging skill trends.

In the Skills-OVATE project [8], Cedefop offered detailed information on the jobs and skills employers demand grouped by regions and sectors. The data were derived from online job postings in 32 European countries, and it is powered by Cedefop’s and Eurostat’s joint work in the context of the Web Intelligence Hub. Skills-OVATE is based on the Tableau Public.

In their report on a Business Intelligence platform, Carnevale et al. (2014) [9] analyzed the growing use of online job advertisements for understanding labor market dynamics, noting their advantages in offering detailed, efficient, and real-time analysis. They highlight the coverage of job ads, skewed towards high-skilled, white-collar positions, and underscore the necessity of using these data alongside traditional labor-market information due to their inherent biases and limitations. They also studied the distribution of job ads within the professional and business services industry to conclude that job ads overrepresent industries that demand high-skilled workers.

Generally, the paper emphasizes the potential of job ads’ data in improving job matching and informing education and training programs, while also cautioning against its exclusive reliance without understanding its limitations.

In a World Bank Group Research project, Brancatelli et al. (2020) [10] introduced an analysis of skill demand in Kosovo with a high level of granularity and precision, such that the results can directly inform policy making. A textual analysis of the job descriptions and job titles was conducted to identify the incidence of skills, education, and experience requirements across industries. For this purpose, they constructed search dictionaries of specific keywords, key phrases, and text patterns to approximate demand for skills, education, and work experience with the occurrence of these in job-portal ads. For the skills analysis, they developed a skill taxonomy that consists of three layers, providing detailed information on skills in demand.

Many other related studies in labor-market analyses use statistical methods, semantic analyses, or vocabulary-based methods in order to extract information from job postings.

Betcherman et al. (2020) [11] investigated the short-term impacts of the COVID-19 lockdown on the Greek labor market based on extracted job-vacancy data from job-search portals that were selected based on Alexa’s web traffic-based rankings. The collected data underwent preprocessing steps such as string cleaning, language detection, elimination of duplicate entries, and standardization of company names and sector affiliations. Advanced machine-learning techniques were applied for the deduplication process, and Natural Language Processing and Named Entity Recognition techniques were used to extract detailed information, including job titles, descriptions, categories, locations, types, contract types, and employer details.

Karakatsanis et al. (2017) [12] proposed a Latent Semantic Indexing (LSI) model that utilizes O*NET occupational descriptions and raw job postings to identify the most demanded occupations in the labor market regardless of the industrial sector and geographical area that is focused on. LSI is based on the idea that there exist implicit semantic connections among the textual data, and that by exploring this latent space, it is possible to gain a deeper understanding of the relationship between words and documents. The
findings reveal that the proposed technique produces much finer and more timely readings of the labor market compared to the common official employment or job-opening statistics and can be directly applied to different labor markets, commercial sectors, and geographical regions.

Sibarani et al. (2017) [13] described an ontology-based information extraction method using a domain vocabulary called Skills and Recruitment Ontology (SARO), which can identify data science skills in job postings, and the performance of the automatic extraction method is compared to its manual (human) equivalent, achieving a satisfactory F-measure of 79–83%. The article also presents a proof-of-concept study that validates the value and potential of the extracted demand data to identify skill-demand composition and trends.

Boselli et al. (2018) [14] explored the prospects that arise from the evaluation of online job postings for labor-market stakeholders, providing a strategic edge over the conventional methods of survey analysis through the facilitation of quicker, evidence-driven decision-making practices. A machine-learning algorithm was designed to categorize vast numbers of internet job advertisements into a standard taxonomy of occupations. The authors shared the practical benefits and insights gained from their approach, highlighting the algorithm’s success in improving activities related to labor-market insights.

Marrara et al. (2017) [15] introduced a method for identifying new professions not yet included in the ISCO standard taxonomy (International Standard Classification of Occupations) [16], using machine learning to classify online job postings. Their work highlights two main contributions: aiding labor-market experts in recognizing emerging occupations for ISCO taxonomy updates and employing language models to find occupations that are similar in their required skills and competencies. The approach, tested on English job vacancies, showed promising outcomes.

Kim and Angnakoon (2016) [17] conducted a comprehensive study assessing the methodologies of research using job advertisements, particularly within the Library and Information Science (LIS) domain. Their study reviewed how job ads have evolved as valuable data sources for understanding labor-market trends, qualifications, and skills required in the LIS field. The research focused on various aspects, including data collection, analysis techniques, and the representativeness of job-ad data.

Karakatsannis et al. (2017) [12] proposed a data-mining approach to identify the most sought-after occupations in the labor market by developing a Latent Semantic Indexing (LSI) model. This model matches job ads from the web with descriptions in the O*NET database, a comprehensive source of occupational requirements. The authors’ research demonstrated the effectiveness of this method for revealing employment trends across different industries and locations, identifying job clusters, and examining changes in occupational demand over time, offering significant insights for job seekers, employers, policymakers, and investors.

Bäck et al. (2021) [18] utilized the job-ad data of a major Finnish labor-market platform to investigate the emergence of AI-related jobs. They created a domain-specific three-tiered vocabulary list of AI-related skills and applied it to the job data to identify the relatedness spectrum of ads to AI and evaluate the usefulness of job advertisements in monitoring technology adoption in companies and the public sector.

During the pandemic, there were several studies that analyzed online job-posting data to gain insights into the impact of the COVID-19 crisis on the labor demand. One such study was by Bamieh et al. (2020) [19], who investigated the impact of the COVID-19 crisis on labor demand in Austria by analyzing job-board data. The authors used data from e-Jobroom, a major Austrian job board, to compare the number of job advertisements posted before and after the outbreak of the pandemic. They also examined the distribution of job vacancies across different sectors and occupational groups. Their paper concluded by highlighting the usefulness of job-board data for monitoring changes in labor demand during the pandemic and the potential implications for policymakers and job seekers.

Table 1 summarizes the focus areas and key findings of the above research work.
Table 1. Literature review summary.

<table>
<thead>
<tr>
<th>Focus Area</th>
<th>Key Findings</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEDEFOP’s contributions</td>
<td>CEDEFOP’s work on labor-market analysis, focusing on VET systems and the impact of economic and social megatrends on skill demands and mismatches. In the Skills-OVATE project, CEDEFOP built a Business Intelligence platform to provide EU detailed information on the jobs and skills employers demand, grouped by regions and sectors.</td>
<td>CEDEFOP [6–8].</td>
</tr>
<tr>
<td>Online Job-Posting Analysis</td>
<td>An analysis of online job-posting data utilizing text-mining and data-mining approaches leads to an understanding of labor-market dynamics and provides real-time insights into job trends and skill demands.</td>
<td>Carnevale et al. (2014) [9]; Brancatelli et al. (2020) [10]; Karakatsannis et al. (2017) [12]; Kim and Angnakoon (2016) [17].</td>
</tr>
<tr>
<td>Machine-learning Techniques</td>
<td>Use of machine-learning, Natural Language Processing, and Named Entity Recognition techniques may be used for labor-information extraction to classify job ads on standard occupation taxonomies and identify emerging occupations and skills, improving labor-market insights.</td>
<td>Boselli et al. (2018) [14]; Marrara et al. (2017) [15].</td>
</tr>
<tr>
<td>Ontology-based information extraction</td>
<td>Application of ontology-based methods and domain-specific vocabularies for extracting data science skills from job postings, showing improved performance over manual methods.</td>
<td>Sibarani et al. (2017) [13].</td>
</tr>
<tr>
<td>Emerging technologies in job-ads Analyses</td>
<td>Investigation of the emergence of AI-related jobs and technology adoption using job-ad data, highlighting the relevance of these ads in monitoring labor-market trends.</td>
<td>Bäck et al. (2021) [18].</td>
</tr>
</tbody>
</table>

3. Methodology

The proposed methodology in the context of gathering online job postings and information extraction refers to a comprehensive approach that encompasses all stages of the process, from initial data collection and data cleansing to the final extraction of relevant information as presented in Figure 1.
Figure 1. Methodology.

As shown in Figure 1, the first step of the methodology was to select reputable and valid sources for our data and then build the mechanism and database structure for data gathering and storing. Once these postings were stored, the methodology includes data cleansing and preparation, the normalization of certain fields, and a deduplication procedure in order to ensure the dataset’s accuracy and quality.

Following the preprocessing steps, the process moved to the extraction of metadata, a task that involves capturing key information like job titles, company names, location, salary, and qualifications from the text of the postings.

In the field of labor-market analysis, CEDEFOP uses a range of methodologies [6,7], which include manual work from experts and costly surveys. The current proposed methodology automates some basic steps of the whole procedure, from online job postings’ raw data to labor-market intelligence. Specifically, in the data-preprocessing phase, we utilized a sequential process that, in addition to CEDEFOP’s methodology, introduces a machine-learning deduplication method and missing values’ handling process. In the information-extraction phase, we presented a skill-extraction use case that utilized an advanced data-driven skills taxonomy instead of manual updates from labor experts.

3.1. The Challenges

In the current work, we introduce a comprehensive methodology for analyzing online job postings to extract valuable labor-market intelligence. There are several challenges that we encountered during this process, each of which has significant implications for the accuracy and reliability of the analysis. Below, we provide a discussion on these.

1. Unbiased and Labor Market-Representative Data:

One of the fundamental challenges is ensuring that the data collected are representative of the labor market at the time of analysis. Online job postings can be biased toward specific
sectors or types of jobs, which may not reflect the broader labor-market trends. To mitigate this, our methodology includes a data-collection process from multiple reputable sources, ensuring diversity and comprehensiveness in the dataset. We emphasize the importance of selecting data sources that capture a wide range of industries and job categories to provide a representative view of the labor market.

2. Noisy or Irrelevant Data:

   Job postings often contain unstructured text with irrelevant information that is not useful for our analysis. This noise can obscure valuable insights and complicate the data-processing steps. Our approach involves a multi-step data-cleansing and preparation process, where irrelevant data are systematically filtered out. We employ techniques such as Natural Language Processing (NLP) to remove unnecessary text and focus on the information that directly pertains to the labor-market analysis.

3. Language and Translation Issues:

   Given that online job postings may be available in multiple languages, linguistic variations pose a significant challenge. Different languages may use different conventions for job titles or descriptions, complicating the classification and extraction of relevant information. Our methodology addresses this by implementing language detection and translation procedures, alongside data quality testing, ensuring that the data can be analyzed consistently across different linguistic contexts.

4. Handling Missing Values:

   Incomplete or missing information in job postings can affect the quality and reliability of the analysis. For instance, missing job titles or skill requirements can lead to inaccurate conclusions about labor-market trends. To counter this, we utilize advanced NLP techniques to infer missing information where possible, ensuring that the dataset remains as complete as possible, without introducing significant biases.

5. Variability in Data:

   The variability in how job titles, company names, locations, and skills are expressed across different postings is another challenge. Such variability can lead to difficulties in normalizing the data for analysis. We apply entity normalization techniques to standardize these variables, allowing for a more accurate cross-comparison and aggregation of data across different job postings.

6. Classification of Ambiguous Data:

   Online job postings often contain ambiguous or poorly defined information, making it difficult to classify data accurately. For example, job titles may not always align neatly with standardized occupational categories. Our methodology uses text-mining and NLP techniques to classify job postings into well-defined categories, such as occupation, industry, and skills, thus improving the accuracy of labor-market insights.

7. Recruiters’ Job Postings:

   A specific challenge arises with job postings from recruiting agencies, where the final employer’s industry might not be explicitly mentioned. This makes it difficult to classify the posting accurately within a specific industry. We address this by applying text-mining techniques to the job-description field, extracting industry information when available and using contextual clues to infer the likely industry classification.

8. Changing Labor-Market Dynamics:

   The labor market is dynamic, with new occupations, skills, and industries emerging over time. Capturing these shifts in real time is crucial for an accurate labor-market analysis. Our methodology incorporates machine-learning techniques that are data-driven, enabling the detection and classification of emerging trends in the labor market. This adaptability ensures that the analysis remains relevant even as the labor market evolves.
The key challenges of our methodology are outlined in Table 2 below.

**Table 2. The challenges.**

<table>
<thead>
<tr>
<th>The Challenges</th>
<th>Description</th>
<th>Suggested Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased and labor market-representative data</td>
<td>The labor data must be diverse and reflect the labor market at the time of the analysis</td>
<td>Online job-posting data are collected from multiple sources. (Section 3.2.1. Data-Source Selection)</td>
</tr>
<tr>
<td>Noisy or irrelevant data</td>
<td>Many job postings, especially in the free-text fields, contain data that are not useful for our data-processing steps and must be removed.</td>
<td>Sequential steps for data cleansing and preparation. (Section 3.3.1. Data Cleansing and Preparation)</td>
</tr>
<tr>
<td>Language and translation issues</td>
<td>Online job-posting data may be available in multiple languages, which can pose challenges for analysis. For example, different languages may have different conventions for job titles or descriptions, making it difficult to classify or extract information from the data.</td>
<td>Data quality-assurance tests (Section 3.3.1. Data Cleansing and Preparation)</td>
</tr>
<tr>
<td>Missing values</td>
<td>Online job-posting data may contain missing or incomplete information, which can affect the quality and reliability of the analysis</td>
<td>NLP techniques to handle missing values. (Section 3.3.4. Missing Values Handling)</td>
</tr>
<tr>
<td>Variability in the data</td>
<td>Job titles, company names, locations, and skills may appear in many different words and expressions but with the same context.</td>
<td>Entity normalization is the process of standardizing entities to a common format in a job-posting analysis. Normalizing entities can help to reduce the variability in the data and make it easier to merge and analyze job postings across different sources. (Section 3.3.2. Entities Normalization)</td>
</tr>
<tr>
<td>Classification</td>
<td>Online job-posting data may contain unstructured or ambiguous information, such as job titles or descriptions that are difficult to classify.</td>
<td>Text-mining and NLP techniques to classify job postings into well-defined categories, such as occupation, industry, and skills. (Section 3.4.1. Industry Extraction; Section 3.4.2. Occupation Extraction; Section 3.4.3. Skills Extraction; and Section 4. Skill-Extraction Use Case)</td>
</tr>
<tr>
<td>Recruiters’ job postings</td>
<td>Many job postings are posted by recruiting companies and do not provide any information about the final employer, making it difficult to classify the ad in a certain industry.</td>
<td>Text-mining and NLP techniques on description field to extract industry information, if it exists. (Section 3.4.1. Industry Extraction)</td>
</tr>
<tr>
<td>Changing labor-market dynamics</td>
<td>The labor market is continuously changing, with new occupations, skills, and industries emerging over time.</td>
<td>Machine-learning, data-driven methods. (Section 4. Skill-Extraction Use Case)</td>
</tr>
</tbody>
</table>
These challenges underscore the importance of careful data management and analysis when analyzing online job-posting data. It is crucial to ensure that the data are representative, reliable, and comprehensive. Appropriate statistical and text-mining techniques are used to handle missing values, classify job postings, and extract meaningful information.

3.2. Data Gathering

The data-gathering step consists of the data-source selection and the data extraction, which is a continuous procedure.

3.2.1. Data-Source Selection

To ensure an effective analysis of a country’s labor market, the first and most crucial step in the raw data-gathering phase is to carefully select the most reputable and widely used public or private job portals. The selection criteria should include the total number of job postings, the range of job categories, and the quality and validity of the job postings’ content. This process may involve seeking input from labor-market experts, conducting online searches, and using specialized SEO marketing tools to identify the most relevant and reliable job portals for each country of interest. Even if the final aim is to analyze a wider area, such as the EU, online data sources must be selected for each country independently as online job portals contain region or country specific job postings, and they do not have international dimension.

The selection process can be challenging due to the vast number of job portals available, each with varying quality and validity of job postings content. This paper proposes a methodology that combines different criteria to select the data sources. These criteria include (1) local expert suggestions, (2) Google search results, and (3) Online SEO Marketing Tools and are analyzed as follows:

1. Local expert suggestions: Country labor-market experts may come from various professions, such as human-resource professionals, government officials, chamber of commerce officers, labor-market analysts, and labor economists. Conducting interviews with these experts can help identify the most appropriate job portals for each country of interest.

2. Google search results: Google search results are widely considered a valuable and unbiased method for selecting the most reputable and widely used websites of a particular category. This is because Google’s ranking algorithms take into account a website’s content quality, quantity, and analytics, such as visitors and page views. In the proposed methodology, Google search results are important criteria for selecting the best job portals in a country.

   The main concept is to use the Google Trends Tool to find the top results that appear on top search queries of the “Jobs” category for a particular country of interest. To do this, the following filters should be selected in Google Trends: (1) the country of interest, (2) the “Past 12 months” time period, (3) the “Jobs” category, and (4) the “Web Search” option.

3. Online SEO marketing tools: Popular SEO marketing tools such as Similarweb.com, Alexa.com, and Moz.com can be used to measure and map the digital world in a timely and comprehensive way. These tools can help us find the most popular and valuable job portals in each country of interest.

Combining these three criteria provides a clear view of a country’s job portals, which can then be used as the data source for job postings.

3.2.2. Data Extraction

Once the most reputable and widely used online job portals have been selected as data sources, the next phase involves developing scraping tools for collecting raw data and building a database structure for data storage. For each portal, both a data-structure and technology analysis are conducted to locate the fields of interest build the scraping tool.
The tool was developed in Python programming language, utilizing the Scrapy framework for extracting data.

The following data are generally collected for every job posting:

- **Job title**: It is the title of the job posting that usually indicates the occupation of the job.
- **Job description**: It is the main text that analytically describes the job vacancy. The job description may contain valuable information that needs to be extracted, as it usually contains details about the job occupation and the responsibilities, the industry of the employer, the workplace, the requested skills and qualifications, etc.
- **Employer’s name**: The employer’s name indicates the name of the company that is posting the job vacancy. However, in many cases, this field may be empty or “confidential”, or it is a recruiting company that posts on behalf of a client. In such cases, the proposed methodology extracts the employer’s industry from the job-description field.
- **Workplace**: The workplace information provides the location of the job posting, which can be used to analyze the distribution of job opportunities across different regions.
- **Employment type**: Employment type is an important factor in labor-market analysis as it provides information about the type of work arrangement between employers and workers. Employment types can include full-time, part-time, temporary, contract, self-employed, and freelance. Understanding the distribution of employment types can help in identifying labor-market trends. It can also provide insights into the availability of different types of jobs in different industries and regions, and help policymakers develop strategies to support job creation and job security for workers.
- **Education level**: The education level requested in a job posting depends on the specific requirements and qualifications necessary for the position. Some jobs may require a high-school diploma or equivalent, while others may require a bachelor’s or master’s degree in a specific field. Additionally, some jobs may require additional certifications or specialized training.
- **Estimated salary**: The salary information provides an estimate of the salary range for the job posting. This information can be used to analyze the salaries offered for different occupations and to identify the factors that influence salary levels.
- **Date posted/expiration date**: The extracted data usually contain the job-posting date and the expiration date and are crucial for the proposed analysis as they can provide valuable information about the dynamic and the changes through time of the labor market.

A database structure is then built with a separate table for each data source to handle the differences in the posting’s fields. The scraping tool is scheduled to run daily and collect new postings. The job posting URL should be used as the table’s primary key to avoid
storing the same posting. Figure 2 presents the structure of the MySQL table where the raw data are stored.

![Raw data DB structure](image)

**Figure 2.** Raw data DB structure.

It is important to note that online job portals are constantly changing, and new job postings are added and removed regularly. Therefore, the scraping tool must be designed to capture the most up-to-date information available.

3.3. Data Preprocessing

Following the data-gathering phase we continue to the data-preprocessing phase and then the information-extraction phase. In order to handle the data and information that come from these phases, we designed a new DB structure, which is presented in Figure 3. The data which are stored in the information extraction DB are analyzed in the following paragraphs of the current work.

![Information-extraction DB structure](image)

**Figure 3.** Information-extraction DB structure.
3.3.1. Data Cleansing and Preparation

Many job postings, especially in the free text fields, contain data that are not useful for our data-processing steps and must be removed. The data-cleansing and -preparation procedure results in clear, useful data that can be used for information extraction. Python code and the libraries chardet, ftfy, BeautifulSoup, re, googletrans, and nltk were used to apply data cleansing and preprocessing techniques. These techniques include fixing encoding issues, removing HTML tags and URLs, replacing noise data with special tokens, translating text to English, standardizing capitalization, and removing stop-words and stemming. These steps ensure that the data are clean, consistent, and ready for further analysis. The main actions involved in data-cleansing and -preparation procedure are as follows:

1. Fix encoding problems: Many online job portals use non-English characters in their postings, which can cause issues with data processing if not properly encoded. For example, if a job posting in a language with non-ASCII characters (such as Greek or Chinese) is not encoded properly, the text may appear as a series of unintelligible symbols. Fixing encoding problems involves identifying and correcting these issues to ensure that the data can be properly processed. The chardet library is used to detect the encoding of text and ftfy (fixes text for you) to fix any encoding issues, ensuring all text is properly encoded in UTF-8.

2. HTML tags removal: Online job postings often contain HTML tags that are used to format the text, such as bold or italicized text. These tags are not useful for our data-processing steps and must be removed to extract only the relevant text. “BeautifulSoup” is used to parse HTML content and remove all HTML tags, retaining only the text content.

3. URLs removal: Some online job postings contain links to external websites that are not relevant to our analysis. These links can be removed to reduce noise in the data and ensure that only relevant information is extracted. At this point, we used regular expressions to identify and remove URLs from the text.

4. Remove noise data: Replace numbers, addresses, phone numbers, currency symbols, etc., with special tokens. Regular expressions are used to replace numbers, addresses, phone numbers, and currency symbols with special tokens.

5. Translate to English: Many online job postings are written in languages other than English. Translating these postings into English may be necessary to ensure consistency and ease of analysis. Automated translation tools such as Google Translate API can be used for this purpose, but it is important to note that these tools may not always produce accurate translations.

6. Capitalize all fields: Standardizing the capitalization of all fields in the job postings can make the data easier to read and analyze. This involves converting all text to uppercase or lowercase letters, depending on the desired format, using Python string methods.

7. Stop-words removal and stemming: Stop-words are common words that do not carry much meaning, such as “the” and “of”. These words can be removed to reduce noise in the data and improve the accuracy of the analysis. Stemming involves reducing words to their base form, such as converting “running” to “run”. This helps to reduce the dimensionality of the data and makes the data easier to analyze. NLTK library was used for the above tasks.

The data-cleansing and -preparation procedure is highlighted in Figure 4:
5. Translate to English: Many online job postings are written in languages other than English. Translating these postings into English may be necessary to ensure consistency and ease of analysis. Automated translation tools such as Google Translate API can be used for this purpose, but it is important to note that these tools may not always produce accurate translations.

6. Capitalize all fields: Standardizing the capitalization of all fields in the job postings can make the data easier to read and analyze. This involves converting all text to uppercase or lowercase letters, depending on the desired format, using Python string methods.

7. Stop-words removal and stemming: Stop-words are common words that do not carry much meaning, such as "the" and "of". These words can be removed to reduce noise in the data and improve the accuracy of the analysis. Stemming involves reducing words to their base form, such as converting "running" to "run". This helps to reduce the dimensionality of the data and makes the data easier to analyze. NLTK library was used for the above tasks.

The data-cleansing and -preparation procedure is highlighted in Figure 4:

- Employer name: Normalizing the employer’s name is a challenging task, as it may vary between different portals. The first step involves removing punctuation and all non-alphabetic characters (other than spaces). Next, a list of “stop words” in companies’ names should be compiled. These “stop words” include the company’s legal entity type (e.g., AE, IKE, OE, and ΕΠΕ for Greece), which can easily be found in a country’s list of legal forms; and frequently used words, such as “Company”, “Corporation”, and “Group”.

- Workplace: Workplaces should be normalized to the standard territory name of the NUTS Taxonomy to obtain accurate workplace statistics in the Information Extraction phase. The preferred level of information is NUTS3; however, higher NUTS levels are accepted if NUTS3 data are not available.

- Education level: Normalizing the various education levels in the dataset is essential for accurate analysis. The International Standard Classification of Education (ISCED) [20] is a commonly used standard classification for this purpose.

- Employment type: Employment type is a crucial piece of information in a job posting, and it should be classified using the International Labor Organization. Classification of status in employment standards [21].

- Salary information: In entities normalization procedure the salary field is cleared from text, it is converted to decimal, and the final value is either the given range (min–max

Figure 4. Data-cleaning and -preparation procedure.

It is important to note that the above actions must be taken carefully, as the removal of relevant data could result in a loss of valuable insights. Additionally, some data-cleansing procedures may require manual intervention, as automated techniques may not always produce accurate results.

3.3.2. Entities Normalization

The normalization of certain fields is an important step in the proposed methodology in order to extract valid statistics. The normalization procedure is applied to several fields, including employer name, workplace, education level, employment type, salary information, and dates, and ensures that the data are consistent and can be analyzed effectively in the next phases.

- Employer name: Normalizing the employer’s name is a challenging task, as it may vary between different portals. The first step involves removing punctuation and all non-alphabetic characters (other than spaces). Next, a list of “stop words” in companies’ names should be compiled. These “stop words” include the company’s legal entity type (e.g., AE, IKE, OE, and ΕΠΕ for Greece), which can easily be found in a country’s list of legal forms; and frequently used words, such as “Company”, “Corporation”, and “Group”.

- Workplace: Workplaces should be normalized to the standard territory name of the NUTS Taxonomy to obtain accurate workplace statistics in the Information Extraction phase. The preferred level of information is NUTS3; however, higher NUTS levels are accepted if NUTS3 data are not available.

- Education level: Normalizing the various education levels in the dataset is essential for accurate analysis. The International Standard Classification of Education (ISCED) [20] is a commonly used standard classification for this purpose.

- Employment type: Employment type is a crucial piece of information in a job posting, and it should be classified using the International Labor Organization. Classification of status in employment standards [21].

- Salary information: In entities normalization procedure the salary field is cleared from text, it is converted to decimal, and the final value is either the given range (min–max...
values) or one value which represents the estimated salary. An extra field should be added to indicate whether we refer to monthly or annual salary.

- Date fields: These fields are converted to date type and are normalized based on a standard date format.

### 3.3.3. Deduplication

As mentioned in the “Data Sources Selection” paragraph, the proposed methodology uses data from multiple sources, and this practice can lead to duplicate data. Many job vacancies are posted in more than one portal and even reposted if the vacancy is not filled within a certain time. Deduplication is a critical step in the data-preprocessing phase, especially when dealing with job-posting data. Duplicate job postings can create inaccuracies in statistical analysis and can skew results. In addition, such duplications can lead to redundancy in the data, which can lead to unnecessary computational overhead and storage issues.

Zhao et al. (2021) [22] introduced a Framework for Duplicate Detection from Online Job Postings. They conducted a comparative study and experimental evaluation of 24 methods and compared their performance with a baseline approach. The experiment reveals that the top two methods for duplicate detection are overlap with skip-gram (OS) and overlap with n-gram (OG), followed by TFIDF-cosine with n-gram (TCG) and TFIDF-cosine with skip-gram (TCS).

In the current work, the deduplication procedure was based on the use of the dedupeio, a python library that uses machine learning to perform fuzzy matching, deduplication, and entity resolution quickly on structured data. Dedupe uses cutting-edge research in machine learning to learn the best way to find similar rows in a dataset, quickly and accurately identifying matches that are considered duplicates. The main steps of the dedupe are as follows:

1. **Training phase:** Users provide a sample of matched and unmatched record pairs. Dedupe uses this sample to train a model, learning how to distinguish between duplicates and non-duplicates.
2. **Blocking:** To improve efficiency, dedupe employs a blocking technique that partitions data into smaller blocks based on certain criteria, reducing the number of comparisons needed.
3. **Prediction:** Once trained, the model predicts the likelihood of pairs of records being duplicates, allowing for automated deduplication and entity resolution.

Overall, deduplication is an essential step in ensuring the accuracy and reliability of a job-posting data analysis. By removing duplicates, analysts can gain a better understanding of the labor market and make informed decisions based on the data.

In the job postings deduplication process, we have to decide what is the time period, concerning the publication dates, where two job postings with the same data (employer, job title, job description, workplace) refer to different job vacancies. Brancatelli et al. (2020) [10], in their work for the World Bank, suggest that job postings with the same data that are posted after 30 days refer to different job vacancies and are not duplicates.

### 3.3.4. Missing Values Handling

Handling missing values is a crucial task in job postings analysis to ensure accurate and complete results. Missing values may occur due to incomplete data or missing fields in job postings, which can impact the analysis.

Peng et al. (2023) [23] presented a systematic and theoretical discussion on the missing values problem in big-data environments, proposed a Monte Carlo likelihood approach for correcting bias in parameter estimation, and suggested that structured reporting practices for missing values can enhance research validity.

In the proposed methodology, the job title is the most important field as it usually indicates the occupation of the posting. Thus, a separate procedure runs to fill in the missing values in the title field. In the proposed methodology a top-to-bottom approach is
utilized based on the ISCO-08 [16] occupations taxonomy extracted from the ESCO portal (European Skills, Competences, Qualifications, and Occupations) [24]. When the title of a job posting is missing, this information may be found in the “Job Description” field.

The proposed procedure tries to find matching expressions and similarity of an occupation description (ISCO-08 4th level) to the job posting’s description using the TF-IDF algorithm to calculate the similarity between the ISCO occupation description and the job posting’s description. The occupation with the highest score is selected as the posting title.

Other fields that need to be examined for missing values are the workplace, the employment type, and the education level. These data are likely to be missing in many job postings; in most job portals, they are not a required input field but are usually provided in the job posting’s description. In the proposed methodology, these missing values are handled using dictionaries based on the normalization of each field and the standard taxonomy that is classified as we described above. The normalized values of these data, as well as alternative expressions of each value, are used to build a dictionary for each field. Next, text-matching and -similarity techniques are used to identify the appearance of these terms in the description field. For example, the dictionary for the employment type includes all the official names from the International Classification of Status in Employment, as well as alternative words and expressions of them (e.g., (2) Independent workers without employees are mostly described as freelancers or contractors).

3.4. Information Extraction

Job postings can contain a wealth of information that can be extracted and analyzed for various purposes. Here are some examples of the information that can be extracted from job postings:

- **Job title:** The job title is typically the first piece of information that can be extracted from a job posting. This can provide insights into the type of job, level of seniority, and responsibilities.
- **Job description:** The job description outlines the duties and responsibilities of the role, as well as the skills and qualifications required to perform the job. This information can be used to understand the requirements of the job and to assess whether a candidate is a good fit.
- **Company information:** Job postings may include information about the company, such as its size, industry, location, and mission. This can provide insights into the company culture and values.
- **Salary and benefits:** Some job postings may include information about the salary and benefits package, such as health insurance, retirement plans, and vacation time. This can help candidates evaluate the compensation package and make informed decisions about whether to apply for the job.
- **Required qualifications and skills:** Job postings often list the required qualifications, such as education, experience, and skills. This information can be used to assess whether a candidate meets the minimum requirements for the job.
- **Application instructions:** Job postings may provide instructions on how to apply for the job, such as submitting a resume and cover letter. This information can be used to determine the application process and timeline.
- **Key performance indicators:** Some job postings may list key performance indicators (KPIs) that the candidate will be responsible for achieving. This can provide insights into the goals and objectives of the role.

By analyzing these and other pieces of information from job postings, recruiters, job seekers, and other stakeholders can gain insights into the labor market, the requirements and expectations for specific roles, and the needs of companies and candidates.

3.4.1. Industry Extraction

The industry sector of the employer is a key factor in labor-market analysis. Understanding the industries that are growing and declining can help policymakers, economists,
and businesses make informed decisions about workforce development, training, and investment. Industry information extracted from job postings can be used to identify skills gaps in the labor market, which can inform workforce development programs and education and training programs. In addition, understanding the industries in which employers operate can provide insights into the types of jobs available and the skills required for these jobs, which can be used to guide decision-making and support economic growth. Overall, understanding the industry of the employer is a crucial component of labor-market analysis that can help to guide decision-making and support economic growth. However, there are challenges associated with classifying employers into the appropriate industry category, such as the fact that a company’s main activity may not always be accurately reflected in official records or reported to relevant authorities.

Several types of analyses can be supported by industry information extracted from job postings. For example, industry information can be used to analyze job postings by industry sector, location, and skill requirements. This information can then be used to identify skills gaps in the labor market, to help inform workforce development programs, and to inform education and training programs.

One approach to industry extraction is to use a taxonomy or classification system, such as the ISIC [25] or the NAICS [26]. The proposed methodology relies on the NACE rev 2.0 taxonomy [27] to classify the industry of employers. NACE is the “statistical classification of economic activities in the European Community” and is required by law to be used uniformly throughout all member states of the European Union.

These classification systems can provide a standard way of categorizing job postings by industry sector and can facilitate comparisons across different regions and industries.

Another approach to industry extraction is to use domain-specific knowledge bases or dictionaries. These resources contain industry-specific terms and phrases that can be used to identify and extract industry information from job postings. For example, the O*NET database contains information on occupations and the industries in which they are typically found. This database can be used to map job postings to industries based on the required skills and qualifications.

Determining a company’s industry sector can be a challenging task. A company may have a variety of activities, and its main activity may not always be accurately reflected in the officially submitted. Additionally, changes in economic activity may not be reported to the relevant authorities [28]. While the most reliable source of information on a company’s industry would be the Official Public Business Registry of a country, not all countries have mechanisms to provide this information, or the data provided may not be reliable.

In industry extraction field, Kühnemann et al. (2020) [28] studied the use of domain-specific keywords to classify enterprises by their economic activity, using NACE codes. They compared a knowledge-based approach with flat classification and a two-level hierarchy, using Naïve Bayes and support vector machine models.

Based on the proposed methodology, a combination of different methods is used to classify employers into the appropriate NACE 2.0 code. The first method involves using online business directories to extract economic-activity data. The preferred source is the country’s official business registry to extract the NACE 2.0 code, but private business directories that contain valid and complete business information should also be considered. An automated mechanism is then built to extract the economic activity code and classify the employer’s dataset.

However, not all employers in the dataset may be classified through the above method. In this case, the proposed methodology builds an industry dictionary [29,30]. This technique involves building a dictionary of keywords that usually appear in the employer’s name or inside a job posting’s description, indicating the employer’s economic activity and the job posting’s industry (e.g., hotel, restaurant, IT solutions, casino, betting, real estate company, technical office, etc.). Each dictionary term is associated with a 4-digit NACE 2.0 category. Using industry dictionaries technique is more accurate and effective when focused on a domain-specific industry, such as a tourism dictionary.
Figure 5 presents the steps of the industry extraction procedure of the job postings’ employers.

The final step of dictionary and NLP techniques handles the “Recruiters problem” too. Many job postings are posted from personnel recruiting companies, and the final employer is not mentioned. By using dictionaries and NLP techniques over the job description, we are able to identify the industry that the posting is referring to.

The classified dataset may then be used as a training set for a machine-learning technique to estimate the employer’s economic activity based on the job titles and descriptions they have posted.

3.4.2. Occupation Extraction

Occupation extraction from online job postings is a critical task for understanding labor-market demands and improving job-matching processes. This involves identifying and categorizing job titles and job descriptions from the unstructured text of job postings. Analyzing labor-market data using occupation type can provide insights into trends and challenges in the labor market. For example, it may reveal that certain occupations, such as healthcare professionals or information technology specialists, are in high demand due to changes in technology or demographic shifts, or the recent pandemic of COVID-19.

Albanesi and Kim (2021) [31] conducted an analytic study on the effects of the COVID-19 recession on the US labor market, based on occupations and how they have changed. Albanesi and Kim (2021) [31] conducted an analytic study on the effects of the COVID-19 recession on the US labor market, based on occupations and how they have changed.

Analyzing labor-market data using occupation type can provide insights into trends and challenges in the labor market. For example, it may reveal that certain occupations, such as healthcare professionals or information technology specialists, are in high demand due to changes in technology or demographic shifts, or the recent pandemic of COVID-19. Albanesi and Kim (2021) [31] conducted an analytic study on the effects of the COVID-19 recession on the US labor market, based on occupations and how they have changed.

The occupation of the job posting can also be extracted using text classification algorithms, which classify the job posting based on the occupation it represents. The ESCO portal provides a detailed list of occupations that can be used for this purpose. As described in the previous section, the job title field can be used to extract occupation information.

While occupation and industry extraction from job postings can provide valuable insights into labor-market trends and help inform workforce development programs, there are also potential limitations to these techniques. One limitation is the potential for bias.
in the data, as certain job titles or industry keywords may be over-represented or under-represented in job postings. Another limitation is related to the accuracy of job titles and descriptions, which can vary greatly and lead to misclassification of occupations and industries. Moreover, the section could be further improved by discussing the need for standardized classification systems to improve the accuracy and consistency of occupation and industry extraction.

There have also been studies targeting specific occupation fields. Papoutsoglou et al. (2019) [32] proposed a mapping study that aimed to extract knowledge from online sources related to the software-engineering labor market. The authors conducted a systematic review, analyzing the types of data sources used, the methods employed, and the outcomes obtained. They found that most studies focused on extracting information from job advertisements, social media, and professional networking platforms, and the most-used methods included text-mining and machine-learning techniques. The study also highlighted the potential of online data sources to inform software-engineering workforce planning, labor-market analyses, and talent-acquisition strategies. The authors concluded by emphasizing the need for more research in this area, particularly in terms of standardizing data-collection and -processing methods.

A method used in occupation extraction is keyword matching, which involves searching for predefined occupation-related keywords in the job posting’s description. These keywords can be obtained from various sources, such as occupation taxonomies, job-posting datasets, and domain-specific dictionaries. The occupation-related keywords can be matched to the job posting’s description using various techniques, such as regular expressions, fuzzy-string matching, and machine learning-based classifiers.

In the occupation extraction field, Schierholz et al. (2020) [33] compared seven algorithms for occupation coding, including both automatic- and statistical-learning approaches, and concluded that certain statistical-learning algorithms can achieve superior performance and may be worth implementing in practice, depending on the specific application and available training data. Djumalieva et al. (2018) [34] proposed a methodology for classifying occupations based on skill requirements found in online job postings. The approach utilized semi-supervised machine-learning techniques on a large dataset of UK job postings. While the paper provided initial results and occupational-group descriptions, the main contribution was the methodology for grouping jobs into occupations based on skills.

The proposed methodology uses a combination of NLP techniques, keyword matching, and machine learning-based methods to extract occupation-related information from job postings’ descriptions and to classify each job posting according to the ISCO-08 4-digit code that best describes the occupation. Firstly, the ESCO Occupations Web Service API is utilized, which takes the job posting’s title as input and returns the 4-digit ISCO-08 code. In cases where the title is missing, the job title can be extracted from the job description field using NLP techniques.

Occupation extraction can be a challenging task, as job titles may vary greatly within and across industries. However, with the use of machine learning and Natural Language Processing techniques, accurate occupation extraction can be achieved [14,35]. The accuracy of occupation extraction can be validated and improved through the use of a classified dataset as a training set for machine learning models.

3.4.3. Skills Extraction

By extracting skills from job postings, analysts can gain insights into the skills demanded by employers in the labor market, and such knowledge can help inform policies related to education and training. It can also help job seekers understand the skills they need to acquire in order to be competitive in the labor market. Additionally, skills extraction can help employers identify skill gaps within their organization and make informed decisions about recruitment and training. Overall, skills extraction from job postings can provide valuable information to multiple stakeholders in the labor market.
Skills extraction from job postings can support various types of analyses, such as those that identify emerging skills gaps, those that analyze the demand for specific skills in different industries and occupations, and those that monitor the skill requirements of specific job roles over time.

One of the challenges of skills extraction is the lack of standardization in the way skills are described in job postings. To address this challenge, researchers have proposed various skills taxonomies and ontologies to provide a standardized vocabulary for skills description. Examples of such taxonomies include ESCO (European Skills, Competences, Qualifications and Occupations), O*NET OnLine, and O*NET-SOC Taxonomy. Zhang et al. (2022) [36] introduced SKILLSPAN, a new dataset for skill extraction from job postings. The dataset contains annotated spans for hard and soft skills, and the authors also introduced two domain-adapted BERT models which show an improved performance in relation to the skill and knowledge components. The authors suggest that their approach to skill extraction has the potential to enrich knowledge bases such as ESCO and contribute to providing insights into labor-market dynamics.

Another challenge in skills extraction is that skills taxonomies and ontologies may become outdated quickly in a rapidly changing labor market. There is also potential for bias in the skills listed in job postings and a lack of information about the level of proficiency required for each skill. The above is handled using data-driven methods and taxonomies. Fareri et al. (2021) [37] introduced SkillNER, a data-driven method for automatically extracting soft skills from text using a named entity recognition system trained with a support vector machine on a corpus of over 5000 scientific papers. The system was validated by a team of psychologists and tested in a real-world case study using the job descriptions of ESCO as the textual source, allowing for the detection of communities of job profiles based on their shared soft skills and communities of soft skills based on their shared job profiles. The tool is useful for firms, institutions, and workers in fostering quantitative methods for the study of soft skills.

Overall, skills extraction from job postings is an essential task in the labor-market analysis, providing insights into the skills that are in demand and the specific job roles that require those skills. By using advanced techniques such as NLP, NER, and machine learning, researchers can extract a wide range of skills from job postings, enabling a more accurate and detailed analysis of the labor market.

4. Skill-Extraction Use Case

The proposed methodology for skills extraction in job postings is based on the work of Djumalieva and Sleeman (2018) in “An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts” [38]. In their study, the researchers developed an algorithmic approach to create an open and data-driven skills taxonomy without expert elicitation. This taxonomy is used by the NESTA organization in their Open Jobs Observatory project [39], which focuses on extracting skills from online job adverts. The article describes the development of a methodology to evaluate the quality of surface forms and to develop skill categories based on the data collected from job postings. The authors used quantitative indicators to assess the specificity of each surface form derived from noun phrases to their skill entity.

The process involves the following key steps:

1. Surface form extraction: A skill-detection algorithm utilizing spacy PhraseMatcher class scanned for “surface form” phrases correlated with the 13,000+ ESCO skills, refined to exclude non-representative terms. These noun phrases, or “surface forms”, are potential skill descriptors.

2. Quality assessment: A machine-learning model is employed to predict the quality of each surface form as a skill entity. This model was trained on a manually labeled dataset of high-quality skill surface forms.

3. Skill entity mapping: High-quality surface forms are mapped to skill entities. A skill entity represents a unique skill concept and may be associated with multiple surface forms.
4. Clustering analysis: Unsupervised machine learning is used (specifically, hierarchical clustering) to aggregate the skill entities into coherent categories. This results in a three-level skills hierarchy. The resulting skills taxonomy, as illustrated in Table 3, consists of the following:
- 8 categories at Level 1 (highest level),
- 15 categories at Level 2,
- 41 categories at Level 3 (most granular level).

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transversal skills</td>
<td>General workplace skills</td>
<td>General workplace skills</td>
</tr>
<tr>
<td>Languages</td>
<td>Languages</td>
<td>Languages</td>
</tr>
<tr>
<td>Healthcare, social work, and research</td>
<td>Care and social work</td>
<td>Care and social work</td>
</tr>
<tr>
<td>Scientific research</td>
<td>Scientific research</td>
<td>Medical specialist skills</td>
</tr>
<tr>
<td>Healthcare</td>
<td>Healthcare</td>
<td>Public health administration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Psychology and mental health</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physiotherapy</td>
</tr>
<tr>
<td>Education</td>
<td>Education</td>
<td>Teaching</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Learning support</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Education management</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extracurricular and sports activities</td>
</tr>
<tr>
<td></td>
<td>Communication</td>
<td>Multimedia and product design</td>
</tr>
<tr>
<td>Sales and communication</td>
<td>Customer services and sales</td>
<td>Marketing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Public relations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Customer services</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sales</td>
</tr>
<tr>
<td></td>
<td></td>
<td>International trade</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transport and logistics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Procurement</td>
</tr>
<tr>
<td></td>
<td>Procurement, logistics, and trade</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information and communication technologies</td>
<td>Information and communication technologies</td>
<td>Data analytics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Web and software development</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IT support services</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Security and cybersecurity</td>
</tr>
<tr>
<td>Business administration, finance, and law</td>
<td>Finance and law</td>
<td>Financial services</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accounting</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Law</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tax</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Business and project administration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Office administration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Human resources</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engineering, construction, and maintenance</td>
<td>Manufacturing and engineering</td>
<td>Manufacturing and mechanical engineering</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Electrical engineering</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Civil engineering</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Automotive maintenance and waste management</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Workplace-safety management</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Horticulture, animal husbandry, and environment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Electrical, heating, and ventilation installation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Construction</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Cont.

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food, cleaning,</td>
<td>Food, cleaning,</td>
<td>Food, hospitality,</td>
</tr>
<tr>
<td>and hospitality</td>
<td>and hospitality</td>
<td>and beauty services</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cleaning services</td>
</tr>
</tbody>
</table>

Figure 6 demonstrates the output of our skill-extraction process for two sample job postings. Each row represents a surface form (skill phrase) extracted from the job description. The columns show the following:

- The extracted surface form,
- The predicted skill entity,
- The three levels of skill categorization,
- A quality score indicating the confidence of the match.

To ensure reliability, we implemented a quality threshold. Based on a manual review, we determined that surface forms with a quality score above 0.3 provided the most reliable skill insights. This threshold helps filter out potentially misclassified or low-confidence skill matches. This approach allows us to not only extract explicit skills from job postings but also to categorize them in a structured manner, providing deeper insights into skill demands across different levels of specificity.

In implementing our skills extraction methodology, we encountered several challenges, as outlined in Section 3.1 of this paper. Here is how we addressed these challenges in the context of skill extraction:

1. Missing values: Skills information in job postings can often be incomplete or implicit. We employed NLP techniques to infer skills from job descriptions when they were not explicitly listed. This involved analyzing the context and requirements described in the posting to deduce likely skill requirements.

2. Language and translation issues: While our initial use case focused on English-language job postings, we designed our methodology to be language-agnostic. The use of surface forms and machine learning for quality prediction allows for adaptation to different languages with appropriate training data.

3. Extraction of skills: To extract meaningful skill information from often-noisy job-posting data, we based our method on the method used by the NESTA Open Jobs Observatory project, which used a skill-detection algorithm with spaCy PhraseMatcher to scan for “surface forms” correlated with ESCO skills.

4. Classification: Classifying skills proved challenging due to the variety of ways skills can be described. We addressed this by using the data-driven skills taxonomy developed by Djumalieva and Sleeman (2018) that provides a flexible and comprehensive framework for categorizing skills across different levels of specificity.

5. Changing labor-market dynamics: The rapidly evolving nature of skills in the job market posed a significant challenge. Our data-driven approach, which does not rely solely on predefined skill lists, allows for the identification of emerging skills.

By implementing these solutions, we were able to create an effective skill-extraction methodology that addresses the key challenges in analyzing job-posting data. Our approach not only extracts skills but also provides a quality score for each extracted skill, allowing for a further analysis of skill demands in the labor market. The resulting insights provide valuable information about skill demands, trends, and emerging needs in the labor market, demonstrating the power of transforming raw job-posting data into actionable labor-market intelligence.
5. Tourism Industry in Greece Use Case

The proposed methodology was applied to an analysis of job postings in the tourism industry in Greece. The research focused on the profound effects of the COVID-19 pandemic on Greece’s tourism industry, a pivotal sector for the nation’s economy. Through a
meticulous examination of over 20,000 online job postings, this study sheds light on the changing dynamics of job demand within this critical sector during the pandemic.

The proposed method faced some main challenges that we addressed in Section 3.1 of the current work:

1. Unbiased and labor-market representative data: In order to derive useful and valid insights, we should collect as many data as possible from various sources that represent the labor market in a certain period. We applied our proposed method for data-source selection, and the OJV portals that came from the above procedure were Indeed, Careerjet, Jobfind, and Karriera.

2. Industry extraction: The main challenge our methodology addressed in the tourism industry-analysis use case is the answer to the question, “what is the tourism industry”?

   Our collected raw data comprised approximately 140,000 online job advertisements posted between July 2019 and August 2021. These data were preprocessed according to the proposed methodology. The crucial phase was the industry extraction, as we had to securely annotate each job posting with the right employer industry. In Greece, the official business registry (GEMI) provides limited access to the companies’ data, so we relied on private business directories that contain valid and complete business information. An automated scraping mechanism was built to extract the economic activity code and classify the employer’s dataset.

   However, not all employers in the dataset were classified through the above method. Moreover, many job postings are posted by recruiter companies, and there is not any information about the final employer. For the above cases, we proceeded to the industry dictionary step of the proposed methodology. In our case, we focused on tourism-industry terms to build our dictionary, such as “hotel”, “restaurant”, “bar”, “tourism”, “resort”, “real estate”, etc. The above procedures resulted in an employer-industry annotated dataset of over 85% of the original postings.

3. Tourism-industry employers: In order to choose only the job postings that belong to the tourism industry, we referred to the work of Demunter and Dimitrakopoulou (2013) [40], as it provides a list of NACE rev 2.0 codes associated with the tourism industries. This resulted in the identification of over 20,000 online job advertisements within the tourism industry.

The findings of our research revealed a significant shift in the labor market of Greece’s tourism sector:

- A 7.5% increase in part-time and short-term contract positions was noted, reflecting the industry’s adaptation to the pandemic’s uncertainties;
- There was a notable rise in demand for high-skilled blue-collar jobs, particularly those requiring tertiary education;
- The skill requirements for tourism roles evolved, with a new emphasis on healthcare, information management, and food service administration skills.

This study also classified job roles into four categories: high-skilled white-collar, low-skilled white-collar, high-skilled blue-collar, and low-skilled blue-collar jobs. This classification provided insights into the changing landscape of employment opportunities in the tourism industry during the pandemic.

This study’s findings are crucial for understanding how the pandemic reshaped the Greek tourism labor market. The increase in part-time and short-term contracts reflects the sector’s response to the uncertain environment. The rise in high-skilled blue-collar jobs highlights a shift towards more specialized roles, potentially driven by changing consumer preferences and health regulations.

Furthermore, the changing skill requirements, particularly the increased demand for healthcare and information management skills, suggest a long-term evolution in the tourism industry’s operational focus. These insights are invaluable for policymakers, educators, and industry stakeholders, as they navigate the post-pandemic recovery and future resilience of Greece’s tourism sector.
6. Discussion

This research successfully demonstrates how online job postings can be transformed into valuable labor-market intelligence. We adopted a comprehensive end-to-end methodology that includes data-gathering, data-preprocessing, and information-extraction steps. We also introduced advanced analysis techniques, like machine learning and NLP in order to gain deep insights into labor-market dynamics. Our approach reassures the credibility of the final insights, as it overcomes some basic issues of the use of LLMs, such as (a) dealing with specific and well-defined problems, (b) working with limited computational resources, and (c) requiring interpretable and explainable results.

Future research should aim to expand the application of this methodology across different geographic regions and industries to validate its effectiveness and versatility. Further refinement of the machine-learning models and NLP techniques could enhance the accuracy and depth of insights extracted from the job postings. The integration of emerging technologies, such as Large Language Models (LLMs) and big-data analytics, may enhance the efficiency and the real-time analysis of labor-market analyses. Future research should explore the potential of these technologies in automating information extraction, improving predictive analytics, and uncovering latent trends within job postings. Lastly, the opportunity to explore the potential biases in online job postings and their implications on labor-market analyses would be a valuable area of investigation.

7. Conclusions

This study presented a comprehensive methodology for transforming online job postings into valuable labor-market intelligence utilizing machine-learning and NLP techniques to extract meaningful insights. Focusing on the Greek tourism industry as a case study, our findings highlight the significant impact of the COVID-19 pandemic on job demands, skill requirements, and employment types within the tourism sector. The skills-extraction use case presented a machine-learning approach that allowed us to extract explicit skills from job postings and categorize them in a structured manner. The data-driven taxonomy that was used may also indicate emerging skill demands in the labor market.

The proposed methodology insights gained from this study provide valuable information for businesses, job seekers, and policymakers, aiding in workforce development and strategic planning. Future research should expand the application of this methodology across different regions and industries, further refining machine-learning models and exploring the integration of LLMs and big-data analytics to enhance labor-market analyses.

Author Contributions: Conceptualization, G.T. and K.C.G.; Methodology, G.T., N.Z., E.M., K.C.G. and P.Z.; Software, N.Z. and P.Z.; Validation, K.C.G.; Formal analysis, E.M.; Resources, N.Z. and E.M.; Data curation, N.Z. and P.Z.; Writing—original draft, N.Z. and P.Z.; Supervision, G.T. and K.C.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

37. Fareri, S.; Melluso, N.; Chiarello, F.; Fantoni, G. SkillNER: Mining and mapping soft skills from any text. Expert Syst. Appl. 2021, 184, 115544. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.