



Creative Fellow 

## Sergej Bogatinoski

### SkillGenius

## Table of Contents

Core Insights and Synthesis .....	4
Origin and Purpose: Why Dive into the World of Job-Skill Matching? .....	5
Significance of the Research Problem .....	6
Research Question .....	7
Project Objectives .....	8
Develop a Personalized Job-Matching Algorithm:.....	8
Incorporate Skill Development Features:.....	8
Create a Community Forum: .....	9
Implement a Feedback Loop for Result Relevancy:.....	9
Ensure Inclusivity:.....	9
Scope and Limitations .....	9
Data Constraints:.....	9
Technological Constraints: .....	10
User Engagement: .....	10
Ethical Considerations:.....	10
Adaptability .....	10
Legal Compliance:.....	10
Journey and Evolution: How I Tackled the Challenge.....	11
Pivot to Semantic Search.....	13

Introduction to the Model and its Functionality .....	13
Development Process.....	14
Challenges and Solutions.....	14
Data Cleaning and Quality Assessment .....	15
ESCO Data Analysis and Processing .....	15
Job Postings Data Analysis and Processing.....	16
Forum Development .....	17
Skill-Skill Relation Component.....	18
Ethical and Legal Considerations.....	19
Learnings and Reflections: What the Journey Taught Me.....	20
Synthesis of Findings & Key Insights.....	20
Understanding the Complexity of Skill-Job Matching .....	20
Evolution in Methodology and Approach.....	20
Importance of User Interaction and Experience .....	21
Data Quality & Its Implications.....	21
Multilingual Challenges & Linguistic Representation.....	21
Practical Applications .....	23
Job Search Platforms .....	23
Career Advising Platforms and Educational Institutions .....	23
Human Resource Management and Recruitment.....	23

Policy Making and Labor Market Analysis .....	23
Potential Impact .....	24
Enhanced Employability and Career Development .....	24
Informed Decision-Making in Education and Training .....	24
Innovative Changes in Recruitment .....	24
Economic Growth and Workforce Optimization .....	24
Enhancement in User-Centric Services.....	24
Inclusive and Equitable Opportunities .....	25
Recommendations for Future Work.....	25
Areas for Further Study .....	25
Suggested Improvements.....	26
Linking back to CIRCE: Reflections and Implications .....	26
Insights into Cultural and Creative Economies.....	27
Understanding Current States and Future Prospects.....	27
Empowering Creative Professions.....	27
Addressing Crises through Cultural and Creative Economies.....	27
The Role of Cultural and Creative Economies in Navigating Crises .....	28
Empowering Creative Communities Through Project Outcomes.....	28
References.....	29

## Core Insights and Synthesis

This project synthesized insights on optimizing job-matching platforms within the cultural and creative economies (CCEs) to address inherent deficiencies in traditional employment paradigms. The central insight discerned was the critical interrelation between advanced technology and CCEs and how this symbiosis could drive innovation, inclusivity, and resilience within these sectors. The project underscored the transformative potential of nuanced, AI-driven methodologies in enhancing job-skills alignment, thus fostering economic growth and social equity in the CCEs.

The traditional paradigms were found to be inefficient, primarily due to their dependence on broad job categories and titles that fail to capture the diverse and specialized skills prevalent in the CCEs. The inefficiencies identified included inadequate personalization, the absence of continuous learning and community building mechanisms, and lack of nuanced job-matching algorithms, all of which detrimentally impacted job placement, career development, lifelong learning, and social inclusivity. Addressing these inefficiencies promised improved employment outcomes, societal equity, and contributions from the CCEs to societal well-being and resilience.

Advanced AI and semantic search were used in resolving these inadequacies, providing more meaningful and relevant matches by understanding the intricacies of language and context inherent in job titles and skills within CCEs. The importance of data integrity was emphasized, with inconsistencies and inaccuracies in job and skill descriptions significantly impacting the reliability of AI models<sup>1</sup>. This highlighted the crucial need for meticulous data assessment and cleaning, especially in the dynamic and diverse fields of the CCEs.

---

<sup>1</sup> An AI model is like a virtual helper that's trained to learn from examples and perform specific tasks, like matching people to jobs. It's designed to mimic the way humans think and learn, allowing it to understand information and make decisions or suggestions. In this case, the AI model analyzes information about jobs and skills, understanding the details and context, to find the best matches. It's like having a super-smart assistant that can sift through tons of information in no time and give you smart suggestions based on what it has learned.

Additionally, the project illuminated the importance of user interaction and experience, leading to the integration of a hybrid system that combines semantic and traditional keyword-based searches to improve alignment between user input and predefined skills in the dataset. This was coupled with the establishment of the Skill-Skill Relation Component, which provides users with actionable pathways for skill development by understanding the interrelations between different skills and competencies.

Ethical considerations, including data privacy and transparency in algorithmic decision-making, surfaced as critical components, necessitating the implementation of ethically sound and legally compliant methodologies, particularly given the sensitive nature of the data. The integration of a community forum and continuous iterations based on user feedback underscored the value of community engagement and adaptability in developing technology solutions. The project also confronted challenges in handling multinational and multilingual datasets, necessitating measures to address linguistic bias and maintain the diversity and authenticity of job market representation across regions.

### **Origin and Purpose: Why Dive into the World of Job-Skill Matching?**

The current landscape of job-seeking and skill matching is a complex ecosystem filled with many challenges, particularly intensified in the CCEs. Traditional paradigms of employment have primarily been restricted to matching job titles or broad categories to potential candidates. This approach has increasingly proven to be inadequate given the intricate and evolving demands of the modern job market (Heine, 2023). The CCEs, for instance, host a workforce comprised of a wide-ranging array of professionals, from freelancers and self-employed individuals to young people at the advent of their careers.

Unlike other sectors, where job roles may fit neatly into predefined categories, the cultural and creative industries often require a blend of specialized skills that defy conventional job titles. For example, a digital artist may require proficiency in not just graphic design, but also project management

and client engagement. Existing job-matching platforms, dependent on rudimentary algorithms, are incapable of capturing such a complex interplay of skills and roles. It is here that the aim of this project, encapsulated in the SkillGenius platform, strives to redress these shortcomings. Leveraging advanced AI algorithms allows for a more nuanced and effective job matching, addressing not just the immediate challenges faced by job seekers and employers, but also contributing to broader objectives of career development and lifelong learning.

Many current platforms act as intermediaries, providing limited opportunities for learning and growth—a critical shortfall particularly for those from underrepresented groups who need extra support to stay competitive. This deficiency is especially significant for these individuals, as they require more accessible resources and opportunities for improvement to keep pace in today's competitive landscape (Eurostat, 2023).

The collaborative nature of the cultural and creative sectors also adds a layer of complexity often neglected by existing job-matching systems. Networking and community engagement are vital for career progression in these industries, yet current job-search platforms lack functionalities that encourage community interaction and peer-to-peer learning.

### **Significance of the Research Problem**

The significance of the research problem lies in its direct impact on the efficacy of the labor market, particularly in the cultural and creative economies. Inefficiencies in job and skill matching not only lead to suboptimal employment outcomes but also contribute to broader issues, such as economic wastage, social inequality, and the stifling of innovation and growth within these specialized sectors.

Firstly, mismatched job placements result in substantial economic costs (Chodnicka, 2020). Employers invest time and resources in the hiring process and further onboarding of employees. When a mismatch occurs, this investment is lost, and the cycle of recruitment must begin anew, consuming additional resources. On the side of the jobseeker, mismatched employment often leads to job

dissatisfaction, reduced productivity, and eventual turnover—all of which are counterproductive to both individual career growth and broader economic progress.

Secondly, inadequate representation and provision for underrepresented groups in existing platforms worsen social inequalities. The lack of targeted features for these populations means that they are often sidelined in favor of more 'traditional' candidates, hindering social mobility and perpetuating cycles of inequality.

Thirdly, the cultural and creative sectors are vital not only for their economic contributions but also for their role in fostering innovation, societal well-being, and cultural diversity. Inefficiencies in job and skill matching within these industries limit their growth potential and societal impact. For instance, when highly specialized roles remain unfilled, projects get delayed, reducing the sector's output and its contributions to society at large.

Lastly, the existing limitations in job-matching platforms hinder opportunities for lifelong learning and career development. In a rapidly evolving job market, continuous skill development is crucial for long-term employability. By not providing avenues for learning and growth, existing systems fail to fulfill a vital function in career longevity and adaptability.

### **Research Question**

The primary research question that drove this project is: "How can advanced AI algorithms be leveraged to create a more effective, inclusive, and personalized job-matching platform, particularly within the cultural and creative economies?" This question encapsulates several sub-questions aimed at addressing the identified gaps and inefficiencies in existing job-matching systems:

- How can AI be utilized to understand the complexities of individual skill sets and align them with specialized job roles, particularly in the cultural and creative sectors?
- What features can be incorporated into a job-matching platform to facilitate lifelong learning and skill development, thus enhancing employability over time?



- How can the platform engage the community in a meaningful way to foster networking, peer-to-peer learning, and a sense of belonging, elements often critical for career growth within the cultural and creative fields?
- How can the system be designed to address the unique needs and challenges faced by underrepresented groups, freelancers, and those at the beginning stages of their careers in the cultural and creative economies?

These sub-questions were formulated to examine various aspects of the central issue, aiming to produce a comprehensive, multi-dimensional solution to the problems identified in the existing landscape of job-seeking and skill matching. Through these research questions, SkillGenius seeks not just to facilitate more accurate job matches but to create a more holistic ecosystem that addresses the diverse needs of job seekers and employers.

### **Project Objectives**

The objectives of the SkillGenius project are manifold, designed to offer a comprehensive solution to the intricate problems identified in the existing job-matching landscape, especially within the cultural and creative economies. The project aims to:

***Develop a Personalized Job-Matching Algorithm:*** Utilize advanced AI techniques to create an intelligent job-matching engine that considers the nuances of individual skills, experiences, and career goals. This algorithm aims to replace the traditional, title-based matching systems with a more dynamic and context-sensitive model, thereby reducing instances of job mismatches.

***Incorporate Skill Development Features:*** Implement features that guide users towards relevant training and educational resources based on the gaps identified between their existing skill set and desired job roles. This aims to foster an environment of lifelong learning, enhancing the employability and adaptability of platform users.

**Create a Community Forum:** Develop a forum-like platform that encourages community interaction, networking, and peer-to-peer advice. The purpose is to incorporate the often-essential elements of networking and community engagement into the job-seeking process, particularly beneficial for the collaborative nature of the cultural and creative sectors.

**Implement a Feedback Loop for Result Relevancy:** Establish a feedback mechanism that rewards users for providing insights on the relevancy of job matches. This feedback loop will not only refine the algorithm's performance but also engage the user community in continuous improvement, making the platform more effective over time.

**Ensure Inclusivity:** Design the platform with specific features targeted towards the unique needs of underrepresented groups, freelancers, and young professionals in the cultural and creative economies. This is to ensure that the system is not just efficient but equitable, leveling the playing field for all users.

By achieving these objectives, SkillGenius aims to surpass the limitations of current job-matching platforms. It aspires to become a comprehensive ecosystem that not only facilitates effective job placements but also enriches the broader landscape of career development, particularly within the cultural and creative economies.

### **Scope and Limitations**

The scope of SkillGenius was conceptualized to be both broad and specific, targeting primarily the European job market while being adaptable to other geographic contexts. During the development of this project, several limitations were encountered:

**Data Constraints:** The insights derived from this project are intrinsically linked to the integrity and quality of two specific datasets: job postings and the ESCO dataset (Directorate-General for Employment, Social Affairs and Inclusion. 2022). Regarding job postings, it is crucial to understand that real-world data is inherently imperfect and working with such data poses a multitude of challenges.

Concurrently, the ESCO dataset, while exceptional in its robustness, is not without its limitations. The shortcomings of the ESCO dataset become evident when confronted with incomplete or sector-biased data, posing risks to the reliability of the generated insights.

***Technological Constraints:*** The AI algorithms utilized for this project required significant computational resources, slowing me down in both the training of the models and their subsequent evaluation.

***User Engagement:*** While the project did not reach the user-engagement phase, I theorized that the utility of the envisioned community forums and feedback loops would be highly dependent on active user participation. This limitation would pose challenges in iteratively improving the platform's algorithms and community-building mechanisms if implemented.

***Ethical Considerations:*** While the platform's design strictly utilizes skills for job-matching, consciously avoiding the incorporation of any biases related to gender, ethnicity, or socio-economic status, it is crucial to acknowledge that the application of such a system in the real-world job market, which is fraught with societal and structural inequalities, might still inadvertently reflect or perpetuate existing biases. Constant vigilance and ongoing adjustments would be necessary to ensure fairness and equality in the job-matching process, mitigating any unintended biases or disparities that might arise in practical implementation.

***Adaptability:*** The rapidly evolving job market, characterized by emerging roles and changing skill requirements, poses challenges to maintaining the long-term efficacy of a platform like SkillGenius.

***Legal Compliance:*** Consideration was given to potential challenges related to the European Union's GDPR regulations (European Parliament and Council of the European Union. 2016), specifically in terms of data privacy and user consent.

Through a proactive examination of these limitations, I have gained invaluable insights into the complexities and challenges inherent in developing a skill-job matching platform. Despite these

constraints, the project demonstrates the potential for a realistic and robust solution to complex problems in the job-matching landscape, especially within the cultural and creative economies.

### **Journey and Evolution: How I Tackled the Challenge**

I began by gathering data from job postings listed on the official EURES website (European Commission, n.d.), which served as a primary source of information in the initial stages. The dataset, featuring jobs labeled with occupation categories based on the ESCO dataset, was central to the advanced algorithm I initially envisioned for skill-job matching.

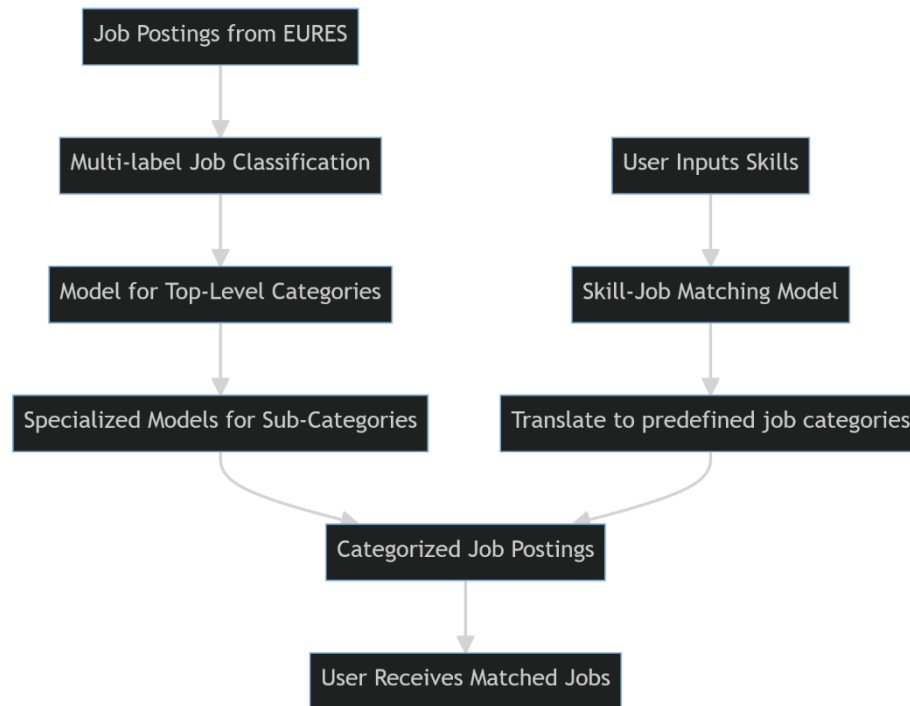
Initially, my strategy for the SkillGenius project involved two major components, both essential for skill-job matching. A high-level overview of how these components work is displayed in Figure 1.

The first component had the objective of classifying job multi-label job classification<sup>2</sup>. To accomplish this, understanding the structure of the ESCO dataset, which was utilized for this task, is crucial. The ESCO dataset is organized in a hierarchical manner, meaning it has multiple levels of job categories, starting from broad types and narrowing down to more specific occupations. Recognizing this hierarchical structure allowed me to conceptualize the classification task as a two-step process. A primary model would categorize job postings into one of the broad, top-level categories. Once this initial categorization was complete, the data would be funneled to specialized models — each trained exclusively on the sub-categories belonging to a specific top-level category — to perform more nuanced classification. This approach was designed to make classifying thousands of occupations more precise and manageable.

---

<sup>2</sup> This term refers to the process of assigning multiple relevant labels or categories to a single job. Imagine you have a job posting; this job may involve various tasks and roles, such as 'data analysis' and 'project management'. Instead of classifying this job under one single category, multi-label classification would allow us to assign multiple, suitable categories to it, reflecting the diverse nature of the tasks involved in the job. This helps in giving a more accurate representation of the job's requirements and responsibilities.

The second major component was planned as a separate mechanism for translating user-input skills into these predefined job categories. The logic behind creating two distinct components was that it would allow for more nuanced and accurate classification of job categories, and secondly, it would facilitate the matching of user-input skills to these classified job postings.



*Figure 1 – High level overview of the first architecture.*

Despite meticulous planning, I ran into two considerable challenges:

The first was the lack of available data to train the mechanism responsible for matching user-input skills to job categories (the second component), which meant that I would need to generate synthetic data, which is far from ideal.

The second obstacle was the underperformance of the job classifiers from the first component. Even training the model for just the top-level categories proved to be challenging. The accuracy of this model plateaued at around 70% to 80%. While this may seem decent for a general classification task, it was unsatisfactory for this project's requirements. This level of accuracy for just the top categories

suggested that training models for the more granular, lower-level classifications—with a whopping total of 3,008 ESCO occupations—would result in even lower accuracy. I experimented with several AI architectures, including BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019) and XLM-RoBERTa (Conneau et al., 2019), but none provided the level of precision I aimed for.

I invested a substantial amount of time in trying to make this intricate system functional. While these efforts may not have led to the desired results, they were crucial to my learning trajectory, and they led me to discover the potential of semantic search (Reimers, N. 2019), a technique that leverages the intricate relationships between job titles, essential skills, and other elements in the ESCO dataset. Transitioning to semantic search was not just a switch in methodology but a fundamental redirection of the project, allowing me to address the original research problem more effectively.

### **Pivot to Semantic Search**

The main advantage of employing a semantic search methodology is that the system could draw on the nuances of language and context to provide more meaningful matches between jobseekers and relevant job categories or skills. The realization that I can use semantic search was by finding out that each job category in ESCO is mapped to essential and optional skills and competencies.

### ***Introduction to the Model and its Functionality***

The model operates by receiving user input in text form. A hybrid search system provided by the open-source database Weaviate (Weaviate, B.V. 2023) combines semantic and traditional keyword-based searches, then suggests several skills from the ESCO dataset that the user can choose from. After selecting one or more skills, users can search for jobs. The user-selected skills, mapped to the ESCO dataset, are sent to the semantic search model, which then finds the most relevant jobs using these user-input skills. This hybrid search system is crucial as it introduces an autocomplete feature that offers skill suggestions based on preliminary user input, enriching user experience and ensuring more precise alignment between user input and the skills outlined in the ESCO dataset.

## ***Development Process***

I initially planned to use the top-level classification model I previously developed as a filter to refine the results generated by the semantic search model. However, extensive testing revealed that using the semantic search model alone yielded more accurate and relevant results.

The process of training a semantic search model involves pairing two sentences, A and B, and providing a similarity score between 0 and 100%. Utilizing the initial job postings data from the EURES website, I was able to map the job occupations in the job postings to the corresponding essential and optional skills required for each job. Additionally, I incorporated the skill-to-skill relation feature within the ESCO dataset. This feature describes which skills are required and optional for a given skill, which can train the model to be able to spot subtle differences, by giving these related skills a lower similarity score than the main skills. After finding all the skills related to a job posting I paired them into two sentences where sentence A contained the skill name and description as described in the ESCO dataset, while sentence B was composed of the job title and description. Essential skills were allocated the highest similarity score of 100%, and the scores were progressively lowered for optional skills and the skills related to these essential and optional skills. Following the experimentation with various AI architectures, I concluded that the multilingual version of E5 (Wang, L. 2022) provided the most precise semantic search results. However, it is important to discuss the specific challenges and constraints encountered during the development and implementation of this AI-driven semantic search system.

## ***Challenges and Solutions***

The foremost issue common to all AI models is crucial: the quality of data, both for the job postings, and for the descriptions of the skills themselves. Data quality posed significant challenges, impacting the accuracy and reliability of the semantic search model. These challenges emanated from inconsistencies within job postings and skills descriptions in the dataset (detailed discussion on data quality issues can be found in the 'Data Cleaning and Quality Assessment' section).

Furthermore, user interaction poses a substantial challenge, especially when users are confronted with articulating their skills in a free-form input, often leading to a struggle to accurately describe the skills they possess. The addition of the hybrid search model was intended to mitigate this issue, enabling users to represent their skills more effectively and accurately. However, it is crucial to acknowledge that this solution is not foolproof and exploring other forms of user experience (UX) enhancements should be a continuous endeavor in the pursuit of an ideal interface and interaction model.

These challenges can be mitigated through the acquisition of user feedback and consultation with domain experts to reassess and refine the skills listed.

### **Data Cleaning and Quality Assessment**

Upon embarking on this journey, one aspect that became increasingly apparent was the importance of data quality, for both the ESCO dataset and the job postings dataset. In both the initial classification model and the subsequent semantic search model, a considerable amount of time was invested in cleaning the data and assessing its quality.

### ***ESCO Data Analysis and Processing***

Although the ESCO dataset was crafted by the European Union and it encompasses 13,890 skills related to 3,008 occupations, its effectiveness is constrained by inherent limitations. The inherent challenge lies in providing detailed, comprehensive insights and relationships for each entry. Furthermore, the continual emergence of new skills and occupations often surpasses the pace at which the dataset can be updated, making it difficult to maintain its relevance and accuracy.

For example, in the dataset, skills related to knowing different programming languages are generalized, with descriptions and associated job roles resembling those of a university lecturer more than those of someone in a practical programming job. As a result, a person searching for a programming job and listing a specific programming language as a skill might receive suggestions that



are skewed towards academic or teaching roles rather than actual programming roles. This was observed during testing where the top results for selecting a programming skill were predominantly lecturer roles.

This kind of oversimplification and lack of specificity is not unique to programming skills; it can potentially occur in other fields as well. The dataset might have similar inaccuracies and generalizations for skills in various domains, possibly leading to mismatches and unmet expectations in those areas too.

### ***Job Postings Data Analysis and Processing***

The issues of data quality were more prevalent in the job postings dataset where I encountered numerous inconsistencies and inaccuracies. These ranged from incorrect language labels to ambiguously defined occupations, incomplete job descriptions, and even grammatical errors. Moreover, I discovered numerous jobs that were miscategorized. These entries inaccurately represented their stated occupations, with the job title and description being either slightly off or completely irrelevant.

Such anomalies and misrepresentations posed a significant risk of misguiding the models during training, thereby compromising the reliability of the matching process. Given the enormity of the dataset and the multitude of languages involved, manual inspection of each entry was impractical.

To address these complexities, a multi-step data cleaning process was undertaken. I started with removing entries lacking substantial information or context about job descriptions and skills.

Subsequently, to tackle the issue of miscategorized jobs, I used GPT-3.5 (OpenAI, 2023) to automatically compare the job titles and descriptions alongside the occupation's descriptions. It returned corrections by updating incorrect labels in the dataset. While GPT-3.5 mostly rendered accurate predictions, there were occasional discrepancies. Manual inspection by domain experts is irreplaceable in ensuring utmost accuracy; however, utilizing GPT-3.5 was a substantial improvement over leaving the dataset with a lot of inaccuracies.

The disproportionate distribution of job postings in multiple European languages posed additional challenges. A significant 85% of the 2.3 million entries were predominantly in German, French, and Dutch. Entries in languages constituting less than 1% of the total dataset were subsequently removed to counter the imbalance.

Surprisingly, I also encountered job postings in languages that were not intended to be in the dataset, such as Japanese. A language detection model, fastText (Bojanowski, P. 2016), was employed to identify and eliminate such anomalies, thereby enhancing data quality.

The meticulous process of data cleaning and quality assessment was not only labor-intensive but was pivotal for the project's success. Attaining high-quality, accurate data was indispensable for training reliable models, forming the backbone of the entire research and development process. The implementation of advanced models like GPT-3.5 and fastText played a crucial role in automating and refining the data cleaning process, bridging the gap between extensive datasets and manual inspection limitations.

### **Forum Development**

In alignment with the overarching aim of this project—to facilitate a more intuitive and holistic job-matching experience—I deemed it essential to incorporate a community forum. The underlying notion was to create a digital space where users could interact, ask questions, share advice, and provide crucial feedback on the AI-driven skill-job matcher. Given that the job market is a constantly evolving entity with numerous intricacies, the forum also serves as a dynamic repository of tips, market trends, and skill development pathways. In doing so, it supplements the AI component, adding a qualitative dimension to the predominantly quantitative analytics.

For the development of the forum, I opted for Svelte (Rich Harris. 2023) and Strapi (Strapi Solutions SAS. 2023) as the foundational technologies. The forum embodies a fusion of technology and social interaction, designed to augment the AI-driven job-skills matching system. It contributes to the

democratization of employment opportunities by fostering a community of well-informed job seekers and thereby enriches the user experience beyond the quantitative mechanics of skill-job matching.

### **Skill-Skill Relation Component**

The Skill-Skill Relation Component emerged as a pivotal facet of the project, informed by the realization that skill-job matching was just one dimension of the complex labyrinth of employability. A more nuanced approach necessitated the understanding of relationships between different skills, knowledges, and competences—both essential and optional—as cataloged in the ESCO dataset. The intention was to offer users not just a list of jobs that matched their skill set but also actionable pathways for skill development that would make them more versatile in the job market.

The ESCO dataset was invaluable in this context, providing structured data on how various skills interrelated. It outlined the essential and optional skills for each skill, knowledge, or competence, effectively laying the groundwork for what would become a recommendation engine for skill improvement. The objective was to leverage this intricate web of relations to provide users with suggestions on complementary skills that could enhance their employability.

Given the intricate nature of skill relationships, the preliminary approach involved employing language models to interpret the data. Utilizing GPT-3.5 seemed like a suitable way forward, as it provided the capabilities needed to generate coherent and contextually relevant suggestions.

The challenge was not only in coding the logic but also in presenting the recommendations in an accessible and user-friendly manner. This was crucial, as the efficacy of the Skill-Skill Relation Component would be gauged by its ability to offer actionable insights to users, which would in turn require a UI that was both intuitive and informative. It was clear from the onset that this would be a complex task that would require iterative improvements based on user feedback, something beyond the project's scope.

## Ethical and Legal Considerations

Navigating the ethical and legal landscape was a critical aspect of the project, one that required attention to existing laws, case precedents, and evolving best practices. The project initially entailed getting job data from the EURES website, raising questions regarding the legality of such an approach. To clarify the legal parameters, I referenced existing cases, notably the LinkedIn vs. hiQ Labs case (*HIQ LABS, INC. V. LINKEDIN CORPORATION, 2022*). The verdict established that public data—information accessible without login or registration—could be legally scraped. Based on this, I concentrated my efforts on gathering only publicly available data.

Ethical considerations were not limited to data collection but extended to data utilization as well. Specifically, the project deals with sensitive employment data, implicating not just job postings but also the skills and competencies of individuals seeking jobs. Therefore, great care was taken to anonymize any user-specific data to avoid any potential misuse or breach of privacy. Only skill-related information provided by the user was utilized, ensuring compliance with privacy norms.

Moreover, the multinational nature of the dataset involved the collection and analysis of data in multiple languages. This raised ethical questions surrounding representation and inclusivity, particularly given that a considerable proportion of job postings were for German/Dutch-speaking positions. The challenge was to prevent a linguistic bias while also respecting the authenticity and diversity of the job market across Europe.

The project's Skill-Skill Relation Component presented an additional layer of ethical considerations. Though the aim was to provide actionable skill development recommendations, there existed a responsibility to ensure that these suggestions were based on reliable data and unbiased algorithms. Transparency in algorithmic decision-making was therefore considered imperative, not merely as a compliance measure but as an ethical necessity.

## **Learnings and Reflections: What the Journey Taught Me**

The six-month fellowship at CIRCE has been a profound learning experience, serving as both a mirror and a window—reflecting the complexities inherent in technological innovation while offering a glimpse into the vast potential for societal impact. This section delves into the project's multifaceted dimensions, dissecting key learnings, and encapsulating reflections that extend beyond the project's immediate scope.

### **Synthesis of Findings & Key Insights**

Throughout the course of this project, the iterative process of exploring, designing, developing, and refining mechanisms for skill-job matching provided invaluable insights into the challenges and potentials within the field of employment and skill categorization. This journey was marked by continuous learning and adaptive strategies to align the project closer to its intended goals, contributing to the realization of a nuanced, multilayered understanding of the initial problems and questions posed. Here is a synthesis of the key findings and insights garnered through this endeavor.

#### ***Understanding the Complexity of Skill-Job Matching***

The initial exploration into the skill-job matching underscored the inherent complexity within the ESCO dataset's hierarchical structure. The broad and multifaceted nature of occupations and the intricate web of associated skills highlighted the necessity for a sophisticated, nuanced approach to navigate the myriad classifications and sub-classifications effectively.

#### ***Evolution in Methodology and Approach***

The shift from a classification-based approach to semantic search marked a significant evolution in methodology. It illustrated how harnessing the nuances of language and context could yield more meaningful, accurate matches between job seekers and relevant job categories or skills. This transition not only addressed the original research problem more effectively but also opened new avenues for enhancing the precision and relevance of skill-job matching for the future.

### ***Importance of User Interaction and Experience***

Addressing user interaction has underscored the paramount importance of user-centric designs in the deployment of AI-driven solutions, offering invaluable insights into the multifaceted nature of user engagement and interaction with advanced technological systems. Crafting intuitive and responsive interfaces proved to be as crucial as developing sophisticated algorithms, necessitating a synergistic approach to align technical robustness with user-friendliness.

### ***Data Quality & Its Implications***

The importance of high-quality, accurate data was an ongoing theme throughout the project. The inconsistencies, inaccuracies, and imbalances encountered within the dataset illuminated the crucial role that refined, reliable data plays in training efficient models and delivering meaningful results. It accentuated the need for meticulous data cleaning and quality assessment to mitigate the risk of compromising model reliability.

### ***Multilingual Challenges & Linguistic Representation***

The aspect of multilingual challenges and linguistic representation emerged as significant focal points in the analysis. The inherent multilingual nature of the dataset, which included job postings predominantly in languages such as German, French, and Dutch, posed substantial challenges due to the disproportionate distribution of languages. The larger representation of certain languages posed the risk of introducing linguistic bias and raised ethical concerns about representation and inclusivity within the employment data. The project had to meticulously navigate this representation to uphold the diversity and authenticity of job markets across Europe, ensuring equal representation of various languages and avoiding the overshadowing of minority languages. Addressing these linguistic challenges was critical to maintaining the integrity and inclusiveness of the project, facilitating a balanced representation of job opportunities across different linguistic communities, and ensuring the broad applicability and relevance of the project across the multilingual European job market.

The analysis led to insights into the limitations and biases inherent in existing big language AI models. The learning trajectory underscored that many AI models, primarily designed and optimized for English, often underperform when deployed in other languages. Although there are models claiming to support numerous languages, their efficacy is typically much lower for languages other than English, due to the lack of balanced training data. For instance, after inspecting a commonly used training dataset, Common Crawl 100 (Conneau et al., 2020; Statistical Machine Translation at the University of Edinburgh, 2020; Wenzek et al., 2020), I found out that 57 languages out of 116, together account for only 1% to the total data size, as shown in Figure 2.

This revelation underscores the existing inequalities and highlights the pressing need to develop solutions for the future of multilingual AI. It is not merely about building models with multilingual capabilities but addressing the deeper issue of unequal data availability and representation across languages.

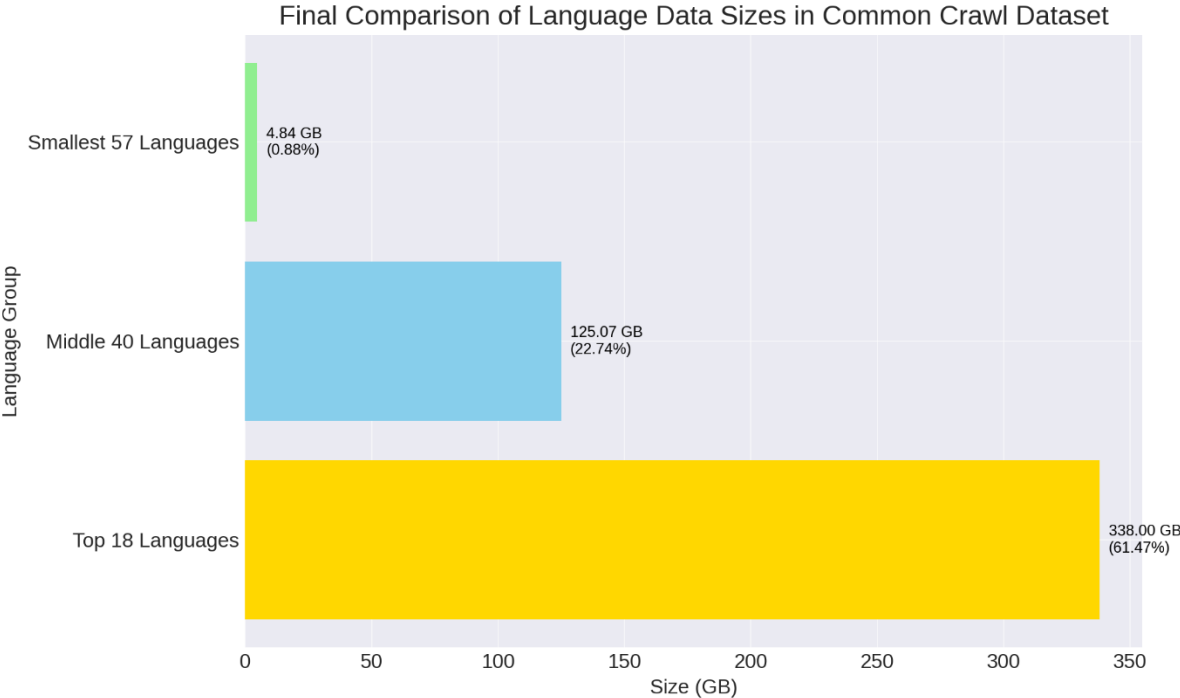


Figure 2 – Bar chart highlighting the data imbalance in the Common Crawl dataset.

## **Practical Applications**

The insights and methodologies developed through this project have versatile applications, particularly in the realms of job matching and career development, offering substantial value to various stakeholders.

### ***Job Search Platforms***

By integrating the project's refined algorithms, job search platforms can enhance their user experience by providing more accurate and personalized job recommendations, enabling users to find suitable employment opportunities more efficiently.

### ***Career Advising Platforms and Educational Institutions***

The refined semantic search methodology and skill-job matching algorithm could be integral for career advising platforms and educational institutions in guiding students and job seekers to relevant career paths and educational programs, based on their individual skills and aspirations.

### ***Human Resource Management and Recruitment***

Companies and HR departments can leverage the advanced job classification and matching techniques developed in this project to refine their recruitment processes. This ensures that the applicants' skills are accurately matched with the job requirements, leading to more informed and effective hiring decisions.

### ***Policy Making and Labor Market Analysis***

The insights can aid policymakers and labor market analysts in identifying trends, skill gaps, and workforce needs, allowing for the formulation of informed policies and strategies to address unemployment and workforce development issues.



## **Potential Impact**

### ***Enhanced Employability and Career Development***

The application of the developed methodologies could lead to more accurate and relevant job matches, empowering individuals to pursue careers that align with their skills and interests, potentially leading to higher job satisfaction and retention rates.

### ***Informed Decision-Making in Education and Training***

Educational institutions and training providers can harness the insights to develop curriculum and training programs that are more aligned with the current market demands, thus preparing the workforce effectively for the future job market.

### ***Innovative Changes in Recruitment***

The integration of advanced job matching techniques can revolutionize recruitment processes, reducing the time and resources spent on finding the right candidate and improving the overall efficiency and effectiveness of recruitment.

### ***Economic Growth and Workforce Optimization***

Improved job-skill matching can lead to optimal utilization of the workforce, contributing to increased productivity and economic growth. Moreover, it can aid in addressing skill shortages and mismatches in the labor market, leading to a more balanced and resilient economy.

### ***Enhancement in User-Centric Services***

By fostering advancements in semantic search and classification models, this project paves the way for the development of more user-centric services and applications, where users receive more accurate and contextually relevant information, enhancing overall user experience across various domains.

### ***Inclusive and Equitable Opportunities***

The application of these insights can contribute to the creation of more inclusive and equitable job markets, where opportunities are accessible to individuals based on their skills and competencies, irrespective of their background, thereby promoting diversity and equality in the workplace.

### **Recommendations for Future Work**

#### ***Areas for Further Study***

**Efficacy of Semantic Search in Different Domains.** A comparison of the efficacy of semantic search models across various domains and industries can be explored to identify any domain-specific considerations or adaptations that may be required.

**Development of Skill-Course Matching Algorithm for the Skill-Skill Relation Component.** To further the capability of the system in aiding individual skill development and acquisition, future work can delve into researching and developing a type of skill-course matching algorithm. This enhancement would allow the system to receive skills derived from the Skill-Skill Relation Component as input and then generate suggestions for relevant courses from popular educational websites as output.

**Comprehensive User Behavior Analysis.** In-depth studies into user interaction patterns and behavior with skill-job matching systems can be conducted to help identify new user needs and preferences, which can inform future system design and development.

**Study of Emerging Job Roles and Skills.** An investigation into the evolving landscape of job roles and requisite skills, with a particular focus on emerging and future trends, can help in maintaining a current and comprehensive knowledge base for the system.

**Evaluation of Long-term Career Trajectories.** Research into long-term career trajectories and the evolution of individual career paths can offer profound insights into career development and enable the development of more personalized and adaptive career advice and job matching.

## ***Suggested Improvements***

**Real-time Data Integration.** Methodologies for seamlessly integrating real-time job market data and trends into the system can be developed to keep the information current and relevant.

**Enhanced Data Preprocessing Techniques.** Refining data preprocessing strategies can be beneficial for handling inconsistencies and inaccuracies more effectively, ensuring the reliability of the models and their outputs.

**Exploration of New Datasets.** Additional datasets beyond ESCO can be investigated to enrich the semantic search and skill-job matching capabilities. Exploring more diverse datasets can contribute to capturing more job roles and skills and mitigating the weaknesses from the dataset enhancing the overall efficacy of the system.

**Expansion to Non-European Job Markets.** To increase the generalizability and applicability of the findings, future endeavors can explore the expansion of methodologies to non-European job markets and analyze their efficacy in diverse cultural and economic contexts.

## **Linking back to CIRCE: Reflections and Implications**

This project, with its deep insights into skill-job matching and the role of advanced technologies, resonates strongly with CIRCE's commitment to practice-oriented science. It has provided actionable insights and highlighted novel ways to leverage technology to bridge gaps between skills and employment opportunities, thus opening avenues to revitalize the cultural and creative industries in Europe. By addressing the complexities of skill-job matching, the project has set the stage for the formulation of policies that are data-driven, efficient, and tailored to the nuanced needs of the creative workforce.

The learnings from the project offer a roadmap to refine policy frameworks that can ensure the optimized flourishing of creative innovations. The advancements in AI-driven skill-job matching contribute to the realization of policies that are more attuned to the evolving needs and dynamics of the

creative and cultural sectors, facilitating the creation of environments where innovation is not merely supported but actively nurtured.

### **Insights into Cultural and Creative Economies**

The project sheds light on the intricate relationship between technological innovation and the cultural and creative sectors, offering reflections on how they can mutually enrich each other. The findings provide a structured perspective on the integration of innovative technologies to enhance the vibrancy and resilience of the cultural and creative economies.

### ***Understanding Current States and Future Prospects***

The nuances unearthed through the project reveal how creative and cultural sectors can navigate through technological transformations, allowing for a more synchronized evolution of both domains. It offers a glimpse into the possibilities of reimagining and reshaping creative professions and industries through informed, innovative applications of technology.

### ***Empowering Creative Professions***

The advancements in skill-job matching technologies uncovered through this project can empower creative professions to redefine their roles and contributions in the societal fabric. It fosters a more profound understanding of the symbiosis between creativity and technology, opening doors for creative communities to effectively address and contribute to solving contemporary global challenges.

### **Addressing Crises through Cultural and Creative Economies**

The role of cultural and creative economies is often underplayed, but it is substantial in confronting and navigating through the multifaceted crises the world is facing today. It is pivotal to look at how the project outcomes can bolster cultural and creative sectors to create transformative impacts in addressing global challenges, and how they, in turn, foster societal well-being and resilience.

### ***The Role of Cultural and Creative Economies in Navigating Crises***

Cultural and creative economies have the intrinsic power to generate innovative solutions and enable societies to adapt and respond to crises more effectively. They are the heartbeat of societal development, providing avenues for expression, dialogue, and understanding that are crucial during times of upheaval. The dynamic and innovative nature of these economies has a transformative potential in addressing socio-economic and environmental challenges and promoting sustainable development.

In contemporary times, marked by rapidly evolving crises such as climate change, socio-political unrest, and pandemics, cultural and creative economies act as catalysts for societal reflection and transformation, fostering a sense of community and shared values, and building the resilience needed to navigate these turbulent times. They create spaces for diverse voices and ideas, contributing to pluralistic and inclusive dialogues that are essential for crisis mitigation and resolution.

### ***Empowering Creative Communities Through Project Outcomes***

The project, with its advanced methodologies and technologies, can significantly empower creative communities. By integrating my findings and innovations, creative sectors can better harness their potential to drive change and contribute to solving pressing global challenges. These advancements can enable artists, creators, and cultural practitioners to expand their reach, amplify their impact, and connect with like-minded individuals and communities, thereby fostering a global network of creative problem solvers.

Moreover, the innovative solutions developed through my project can facilitate the creation of platforms and ecosystems that enable creative communities to collaborate, share knowledge, and access resources, thus enhancing their capacity to generate meaningful and sustainable contributions to societal well-being and development.

## References

- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). *Enriching Word Vectors with Subword Information*. arXiv.org. <https://arxiv.org/abs/1607.04606v2>
- Chodnicka, M. C. (2020, September 30). *How businesses waste their money on recruitment*. LinkedIn. <https://www.linkedin.com/pulse/how-businesses-waste-money-recruitment-magdalena-chodnicka>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.747>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv.org. <https://arxiv.org/abs/1810.04805>
- Directorate-General for Employment, Social Affairs and Inclusion. (2022). *European Skills, Competences, Qualifications and Occupations (ESCO) (V1.1.1)* [Dataset]. <https://esco.ec.europa.eu/en/news/esco-v111-live>
- European Commission. (n.d.). *Find a job in Europe*. Europa. Retrieved September 24, 2023, from [https://ec.europa.eu/eures/portal/jv-se/home?lang=en&pageCode=find\\_a\\_job](https://ec.europa.eu/eures/portal/jv-se/home?lang=en&pageCode=find_a_job)
- European Parliament and Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504>
- Eurostat. (2023). *Migrant integration statistics - education*. <https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/37413.pdf>

Heine, A. H. (2023, July 14). *How long does it take to find a job?* Indeed.

<https://www.indeed.com/career-advice/finding-a-job/how-long-does-it-take-to-find-a-job>

HIQ LABS, INC. v. LINKEDIN CORPORATION, No. 17-16783 (9th Cir. 2022). (2022, April 18). Justia Law.

<https://law.justia.com/cases/federal/appellate-courts/ca9/17-16783/17-16783-2022-04-18.html>

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTa: A Robustly optimized BERT Pretraining Approach*. arXiv.org. <https://arxiv.org/abs/1907.11692>

OpenAI. (2023). *GPT-3.5-turbo* [Large language model].

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*.

arXiv.org. <https://arxiv.org/abs/1908.10084v1>

Rich Harris. (2023). *Svelte* (V3.54) [Software]. <https://svelte.dev/>

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv.org. <https://arxiv.org/abs/1910.01108>

Statistical Machine Translation at the University of Edinburgh. (2020). *CC-100* [Dataset]. European Language Resources Association. <https://data.statmt.org/cc-100/>

Strapi Solutions SAS. (2023). *Strapi* (V4.12.1) [Software]. <https://strapi.io/>

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., ...Wei, F. (2022). *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. arXiv.org. <https://arxiv.org/abs/2212.03533v1>

Weaviate, B.V. (2023). *Weaviate* (V3.4) [Software]. <https://weaviate.io>

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2020).

*CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data*. European Language Resources Association. <https://www.aclweb.org/anthology/2020.lrec-1.494>