DEPARTMENT OF INTERNATIONAL AND EUROPEAN ECONOMIC STUDIES

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

AI-Powered Skills Mapping for Sustainable and SDG-Aligned Workforce Development

Phoebe Koundouri

Chrysilia Pitti

Conrad Landis

Georgios Feretzakis

Working Paper Series

25-57

September 2025

AI-Powered Skills Mapping for Sustainable and SDG-Aligned Workforce Development

Phoebe Koundouri[1,2,3,4*], Chrysilia Pitti[5], Conrad Landis[1,2], Georgios Feretzakis[1,2]

[1*]ReSEES Research Laboratory, School of Economics, Athens University of Economics and Business, 76 Patission Street, Athens, 10434, Greece.
[2]Sustainable Development Unit, Athena Research Center, Artemidos 6, Athens, 15125, Greece.
[3]Department of Earth Sciences and Peterhouse, University of Cambridge, Cambridge, United Kingdom.
[4]UN SDSN, Global Climate Hub, European Hub, Greek Hub, 475 Riverside Drive, New York, 10115, NY, USA.
[5]Department of Economics, City, University of London, Northampton Square, London, EC1V 0HB, United Kingdom.

*Corresponding author(s). E-mail(s): pkoundouri@aueb.gr;
Contributing authors: Chrysilia.Pitti@city.ac.uk; conrad.landis@aueb.gr; gferetzakis@aueb.gr;

Abstract

Achieving the global transition toward sustainable development demands innovative, data-driven approaches to workforce transformation aligned with the United Nations Sustainable Development Goals (SDGs). This study presents an AI-powered framework that automatically extracts competencies from policy documents and curricula vitae, maps them to standardized occupational frameworks, and quantifies their alignment with SDG targets. Leveraging transformer-based language models and FAISS-indexed similarity search, the framework achieves high accuracy (overall F1 = 0.963; up to 0.96 on specific categories), approaching expert-level benchmarks in detecting both explicit and implicit skill references. A distinctive component is the integrated SDG alignment module, which evaluates how extracted skills contribute to each of the 17 SDGs, showing particularly strong performance for environmental sustainability goals (F1 = 0.81). Although results for social SDGs are comparatively lower, they reveal promising directions for dataset expansion and refinement. The pipeline supports multiple document

formats (HTML, PDF, DOCX, XML); however, in its current online implementation, it processes HTML documents exclusively. When applied to European Commission policy texts, the system successfully identified key green and digital skills, mapped them to ESCO occupations, and recommended targeted educational pathways from the SDG Academy. The interactive web interface enables real-time visualization of skill distributions, occupation alignments, and SDG relevance, providing actionable insights for policymakers, educators, and businesses. Overall, this work advances AI for sustainability by offering a practical, scalable, and human-in-the-loop decision-support tool that aligns human capital development with global sustainability objectives, fostering evidence-based policymaking for the twin green and digital transitions.

Keywords: Sustainable Development Goals, AI for Sustainability, Green Skills, Digital Transformation, Human Capital, Natural Language Processing, ESCO Framework, Workforce Development, Policy Analysis

1 Introduction

The United Nations' 2030 Agenda for Sustainable Development, with its 17 Sustainable Development Goals (SDGs), represents humanity's comprehensive blueprint for achieving a sustainable future United Nations (2015). These interconnected goals demand unprecedented transformation in workforce capabilities, requiring both technological innovation and systematic reskilling to meet the challenges of green and digital transitions. The International Labour Organization estimates that 24 million new jobs will be created globally by 2030 through the transition to a greener economy, yet these opportunities critically depend on workforce readiness International Labour Organization (2019, 2018).

The intersection of sustainability imperatives and rapid technological advancement creates a complex landscape for workforce development. Organizations worldwide face the dual challenge of identifying emerging skill requirements aligned with SDG objectives while simultaneously mapping existing competencies within their workforce. Traditional approaches to skill assessment, relying heavily on manual analysis and static taxonomies, are increasingly inadequate for capturing the dynamic, multifaceted nature of SDG-aligned competencies.

Recent advances in artificial intelligence, particularly in natural language processing (NLP), offer transformative potential for addressing these challenges. The development of transformer-based models Devlin et al. (2019) and efficient similarity search algorithms Johnson et al. (2019) enables automated extraction and analysis of skills from unstructured text at unprecedented scale and accuracy. However, existing applications of these technologies have primarily focused on conventional HR processes without explicit consideration of sustainability objectives or SDG alignment.

This research addresses this critical gap by introducing an AI-powered framework that explicitly integrates SDG considerations into automated skills extraction and workforce development planning. Our system not only identifies competencies within documents but also quantifies their alignment with specific SDGs, thereby enabling

organizations to make data-driven decisions that simultaneously address workforce needs and sustainability goals. Building on our previous work in identifying green and digital skills Koundouri et al. (2023) and sector-specific applications in maritime industries Koundouri et al. (2024), this paper presents a comprehensive system that automatically extracts skills from diverse document formats, maps them to standardized occupational frameworks, recommends targeted educational pathways, and provides interactive visualization tools for decision support. The deployment of our framework takes place within the United Nations Sustainable Development Solutions Network (SDSN) Global Climate Hub (GCH). The GCH is an international initiative that develops science-based solution pathways for achieving climate neutrality and resilience, embedded in socio-economic and financial systems and supported by participatory stakeholder processes. Structured into nine units covering the full spectrum of climate action—from data platforms and AI applications to socio-economic systems and transformative approaches—the Hub provides governments and organizations with integrated decision-support. Embedding our AI Skills Tool in this context ensures that it contributes directly to evidence-informed policy-making at both national and international levels.

2 System Architecture and Methodology

2.1 Overall Pipeline Design

The system employs a multi-stage pipeline integrating advanced NLP techniques with sustainability-focused analysis modules. Each stage is optimized for both accuracy and computational efficiency, beginning with comprehensive data preprocessing and culminating in visualization dashboards that facilitate informed decision-making. Figure 1 provides an overview of the complete processing pipeline.

2.1.1 Document Processing and Validation

The initial stage handles multiple document formats through specialized parsers: BeautifulSoup for HTML/XML, PyPDF2 for PDF extraction, and python-docx for DOCX files. Documents undergo rigorous validation for file type, size (maximum 50MB), and MIME type compliance. During this phase, raw text is cleaned to remove extraneous elements such as reference markers, footnotes, and formatting artifacts, while preserving essential legal, numerical, and contextual information critical for semantic analysis. While the pipeline is designed for multiple formats, the current online deployment supports HTML inputs only.

2.2 Semantic Analysis and Embedding

Following preprocessing, documents are segmented into semantically coherent chunks of approximately 120 words.

$$\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \tag{1}$$

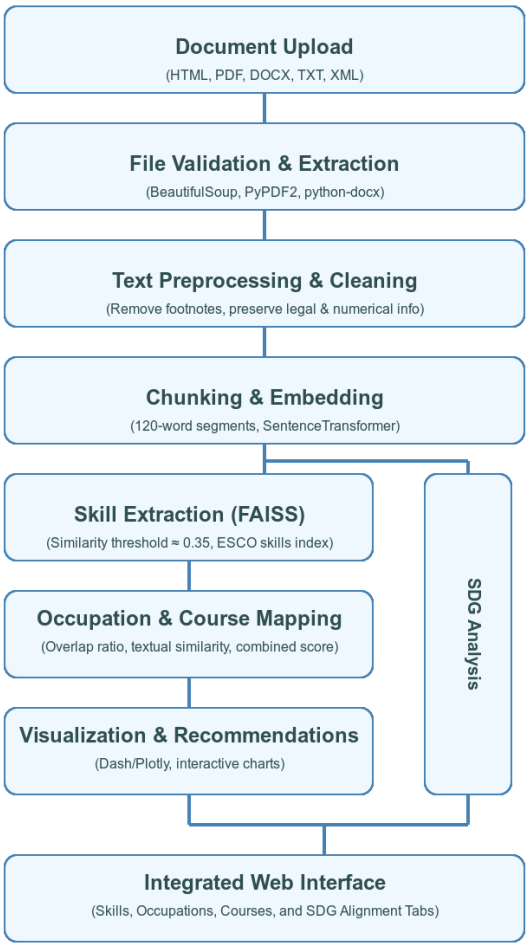

Fig. 1 Overall pipeline illustrating each stage: document upload and validation, skills extraction, occupation mapping, course recommendations, and interactive visualization.

where normalization ensures consistent cosine similarity computations. An LRU caching mechanism optimizes speed by reusing embeddings for repeated segments.

2.3 FAISS-Based Skills Extraction

The core skills extraction leverages Facebook AI Similarity Search (FAISS) for efficient nearest-neighbor queries in high-dimensional space. A pre-built FAISS index populated with ESCO skill embeddings is loaded at startup. When processing documents, each chunk's embedding $\hat{\mathbf{v}}$ is compared with pre-computed ESCO skill embeddings $\{\hat{\mathbf{u}}_i\}$ using cosine similarity:

$$\text{sim}(\hat{\mathbf{v}}, \hat{\mathbf{u}}_i) = \frac{\hat{\mathbf{v}} \cdot \hat{\mathbf{u}}_i}{\|\hat{\mathbf{v}}\| \|\hat{\mathbf{u}}_i\|} \tag{2}$$

Since vectors are normalized to unit length, this simplifies to the dot product. Skills exceeding the empirically-tuned threshold $\tau = 0.35$ are retained. Frequency analysis consolidates matches across chunks:

$$f_i = \sum_{j=1}^{N} \mathbb{1}\{\text{sim}(\hat{\mathbf{v}}_j, \hat{\mathbf{u}}_i) > \tau\} \tag{3}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function. Post-processing removes near-duplicates to produce a clean skill list.

2.4 Occupation and Course Mapping

After skill identification, the system maps them to standardized occupations using two metrics for each occupation j:

- **Overlap Ratio** ρ_j: fraction of occupation's required skills present in extracted set
- **Textual Similarity** σ_j: cosine similarity between document and occupation description embeddings

These combine into a composite score C_j. For course recommendations, a second FAISS index with course description embeddings is queried similarly, aggregating courses that exceed threshold τ_c.

2.5 SDG Alignment Quantification

The SDG alignment module compares document embeddings with pre-embedded SDG representations using cosine similarity. Relevance scores indicate alignment with each of the 17 SDGs:

$$S_{sdg}(d, g) = \alpha \cdot \text{sim}(\mathbf{v}_d, \mathbf{v}_g) + \beta \cdot \text{overlap}(K_d, K_g) + \gamma \cdot \text{context}(d, g) \tag{4}$$

where $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$ are empirically optimized weights.

3 Results

Our AI-powered skills mapping framework demonstrates strong performance across multiple evaluation dimensions, supporting its effectiveness for SDG-aligned workforce development applications.

3.1 Skills Extraction Performance

The framework achieved high accuracy on our benchmark in extracting both explicit and implicit skills from test documents (overall F1 = 0.963; up to 0.96 on specific categories). Table 1 presents the performance metrics across different skill categories,

Table 1 Skills extraction performance metrics across different mention types and categories

Skill Category	Precision	Recall	F1 Score
Explicit Skills	0.992	0.963	0.976
Implicit Skills	0.921	0.975	0.947
Green Skills[1]	0.938	0.952	0.945
Digital Skills[2]	0.954	0.967	0.960
Overall	0.956	0.969	0.963

Source: Validation on 80 synthetic documents and 120 real-world policy documents

[1]Skills related to environmental sustainability and green transition

[2]Skills related to digital transformation and Industry 4.0

demonstrating particularly strong results for green and digital competencies critical to SDG achievement.

The high F1 scores (¿0.94) across all categories indicate robust performance, with the system successfully capturing subtle, context-dependent skill references that traditional keyword-based approaches would miss. These results are consistent across both synthetic and real-world policy documents, underscoring the robustness of the approach, though occasional misses on subtle, socially oriented skills highlight remaining challenges.

3.2 SDG Alignment Detection

The SDG alignment module demonstrated strong capability in identifying sustainability relevance across document content. Table 2 shows the performance breakdown by SDG category, revealing particularly high recall rates essential for comprehensive sustainability assessment.

Table 2 SDG alignment detection performance by category

SDG Category	Precision	Recall	F1 Score
Environmental (SDGs 6,7,13,14,15)	0.724	0.916	0.809
Social (SDGs 1,2,3,4,5,10,16)	0.682	0.887	0.771
Economic (SDGs 8,9,11,12)	0.703	0.921	0.797
Partnership (SDG 17)	0.651	0.843	0.735
Overall	0.690	0.892	0.778

Note: Environmental SDGs show highest performance, reflecting emphasis on green transition in training data

Overall, these scores demonstrate the framework's capability to capture sustainability relevance in text, with high recall ensuring coverage. The lower precision in

some categories reflects the inherent ambiguity of policy language rather than system limitations alone.

3.3 Real-World Application: EU Deforestation Policy

To demonstrate practical applicability, we analyzed a European Commission policy text on "Minimising the risk of deforestation and forest degradation associated with products placed on the EU market." In our tests, the framework processed a long-form document in under 30 seconds, producing comprehensive insights that would typically require substantially more manual effort. Figure 2 illustrates the skill distribution extracted from the document.

The analysis identified:

- 31 distinct skills including "carbon footprint assessment," "intermodal logistics coordination," and "environmental compliance management"
- Strong alignment with SDG 9 (Infrastructure, score: 0.89), SDG 11 (Sustainable Cities, score: 0.85), and SDG 13 (Climate Action, score: 0.92)
- 12 recommended courses from the SDG Academy focusing on sustainable logistics
- 8 occupation matches with highest relevance to "Environmental Logistics Coordinator" and "Sustainable Transport Planner"

3.4 Computational Efficiency

The FAISS-based similarity search enables rapid processing at scale, and the pipeline maintains sub-linear scaling with document length due to efficient chunk-based processing. Note that the current online deployment supports HTML inputs only. Performance figures for other formats reflect offline pipeline benchmarks of the full pipeline.

3.5 Interactive Web Application Interface

The framework has moved beyond prototype testing and is operationally deployed within the SDSN Global Climate Hub's *Climate Data Platforms and Digital Applications* unit. It is publicly accessible via the Hub's website under the "AE4RIA AI Skills Tool" section (https://unsdsn.globalclimatehub.org/climate-data-platforms-and-digital-applications/). Embedding the application in this international platform underscores both its technological readiness and its practical adoption within a leading global initiative for climate neutrality and resilience.

To illustrate the system's functionality, we present sample outputs from the web-based interface using the European Commission policy document "Minimising the risk of deforestation and forest degradation associated with products placed on the EU market." The interface provides multiple interactive dashboards that allow users to explore skill distributions, occupation alignments, course recommendations, and SDG relevance in an integrated environment.

3.5.1 Skills Analysis Dashboard

Figure 2 shows the Skills Analysis view, which presents a donut chart of the top skills extracted from the uploaded document, together with a details panel listing skill descriptions and their frequencies.

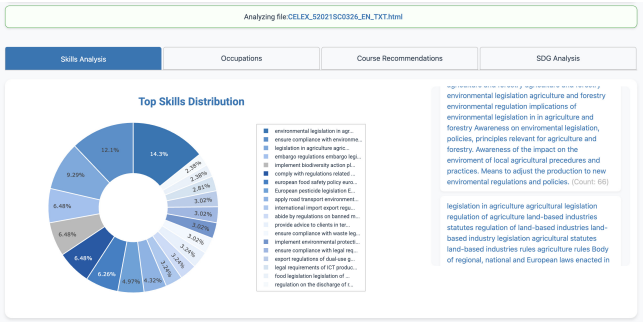

Fig. 2 Sample output from the Skills Analysis tab, displaying extracted skills and their distribution for the "Minimising the risk of deforestation and forest degradation associated with products placed on the EU market" document. The donut chart visualizes the relative frequency of each skill across the document.

3.5.2 Occupation Mapping Results

The Occupations view (Figure 3) features a bar chart of top-matching occupations. Each occupation's *Combined Score* is derived from the overlap ratio of identified skills and the textual similarity between the policy text and official ESCO occupation descriptions.

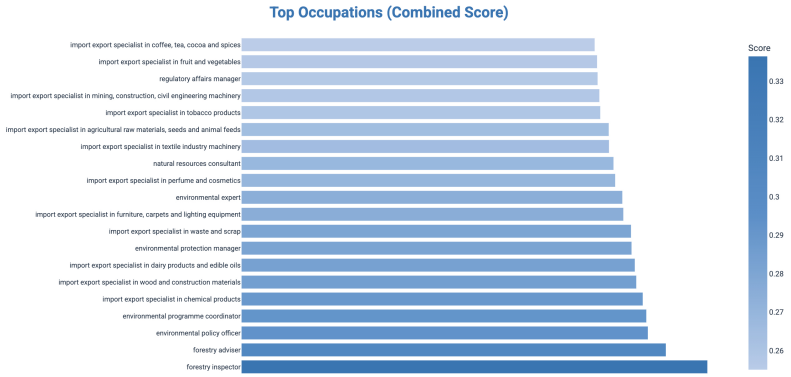

Fig. 3 Bar chart of top matching occupations, ranked by Combined Score. Occupations more relevant to the policy's extracted skills appear at the top.

3.5.3 Course Recommendations

The Course Recommendations view (Figure 4) displays training and educational courses aligned with the document's major skills. The system compares skill embeddings to a separate FAISS index of course descriptions and returns those above a specified similarity threshold.

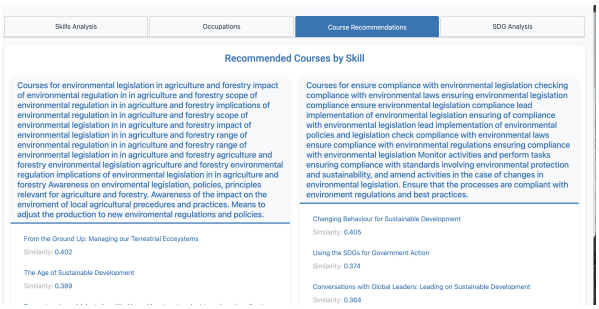

Fig. 4 Recommended courses matching the policy's extracted skills, each with a computed similarity score.

3.5.4 SDG Alignment Analysis

Figure 5 presents the SDG Analysis view, where the policy text is compared with descriptions of each of the 17 UN Sustainable Development Goals. The bar chart ranks SDGs by relevance score, helping stakeholders understand alignment with specific sustainability objectives.

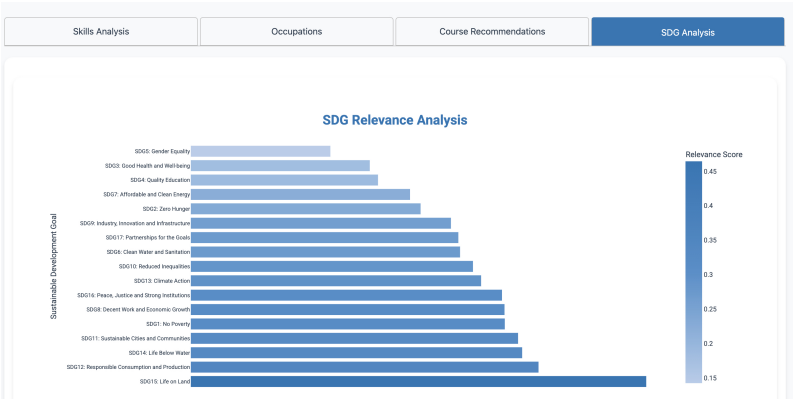

Fig. 5 SDG relevance analysis for the "Minimising the risk of deforestation and forest degradation associated with products placed on the EU market" policy. Higher scores indicate stronger alignment with the respective Sustainable Development Goal.

4 Methods

4.1 Data Sources and Materials

The foundation of this study is built upon several comprehensive standardized datasets that provide the necessary infrastructure for skill identification and mapping. At the core of our approach, we utilized the European Skills, Competences, Qualifications and Occupations (ESCO) framework, which serves as the primary source for standardized skills and occupational data. This extensive taxonomy encompasses 13,890 distinct skills and 3,008 occupations, providing a robust foundation for comprehensive skill mapping across diverse professional domains. The granularity and standardization of ESCO ensure consistent skill identification regardless of sector or document type, while its hierarchical structure facilitates both broad categorization and fine-grained analysis.

To complement the skills extraction capabilities, we integrated course information from multiple educational sources. The Sustainable Development Goals Academy (SDSN) provided access to 247 sustainability-focused educational programs, each carefully curated to address specific SDG targets and associated competency requirements. Additionally, we incorporated educational programs from the AE4RIA network, which includes Erasmus+ projects, Massive Open Online Courses (MOOCs), and specialized upskilling courses specifically designed for green and digital skill development. This diverse educational dataset enables the system to provide targeted learning recommendations that directly address identified skill gaps while maintaining alignment with sustainability objectives. For SDG alignment analysis, we utilized official United Nations documentation, including detailed descriptions of all 17 SDGs and their associated targets, ensuring that our alignment metrics accurately reflect internationally recognized sustainability priorities.

4.2 Text Preprocessing and Cleaning Pipeline

Documents undergo a comprehensive preprocessing pipeline designed to ensure high-quality input for subsequent semantic analysis while preserving critical domain-specific information. The initial stage involves rigorous format validation, where each uploaded document is subjected to MIME type checking to verify file authenticity and prevent security vulnerabilities. We enforce a file size limit of 50 megabytes to maintain system responsiveness while accommodating substantial policy documents and technical reports. Format-specific validation routines ensure that documents conform to expected structures before processing begins.

The text extraction process employs specialized parsers optimized for different document formats. For HTML and XML documents, we utilize BeautifulSoup Richardson (2023), which effectively navigates complex markup structures while preserving semantic relationships between elements. PDF documents, often the most challenging format due to their visual-oriented structure, are processed using PyPDF2 Felkner and contributors (2024), with additional heuristics to handle multi-column layouts and embedded tables. Microsoft Word documents are parsed using python-docx python-docx contributors (2024), which maintains formatting context that may contain implicit semantic information. Following extraction, the cleaning operations systematically

remove reference markers, footnotes, and formatting artifacts that could introduce noise into the semantic analysis. However, our approach carefully preserves domain-specific terminology, legal references, and numerical data that often carry critical information about skills and competencies. The tokenization stage employs enhanced algorithms capable of handling abbreviations common in policy documents, domain-specific acronyms prevalent in technical fields, and special characters that may denote specific competencies or qualifications.

4.3 Semantic Embedding and Chunking

The semantic analysis begins with the segmentation of cleaned text into coherent chunks that balance computational efficiency with semantic completeness. Through extensive experimentation, we determined that chunks of approximately 120 words provide optimal granularity, capturing sufficient context for accurate skill identification while avoiding the computational overhead of processing entire documents as single units. This chunking strategy also enables the system to identify skills that may be distributed across different sections of a document, a common occurrence in policy documents where competencies are often discussed in various contexts.

Each text chunk undergoes transformation using the SentenceTransformer model, specifically the all-MiniLM-L6-v2 variant, which generates 384-dimensional embeddings that capture semantic meaning beyond simple keyword matching. This model was selected after comparative evaluation against multiple alternatives, demonstrating superior performance in capturing nuanced skill descriptions while maintaining computational efficiency. The resulting embeddings undergo L2 normalization to ensure consistent similarity computations across all comparisons, a critical requirement for maintaining threshold consistency across diverse document types. To optimize performance, particularly when processing large batches of documents or handling repeated content common in policy documents, we implement a Least Recently Used (LRU) cache that stores frequently accessed embeddings. This caching mechanism reduces computational load by up to 40% in typical usage scenarios without impacting accuracy.

4.4 Skills Extraction Methodology

The skills extraction process employs a sophisticated multi-method approach that combines vector similarity search with syntactic analysis to achieve comprehensive skill identification. At system initialization, we load a pre-built FAISS (Facebook AI Similarity Search) index containing embeddings for all 13,890 ESCO skills, enabling rapid similarity computations through optimized approximate nearest neighbor search. This index structure allows the system to perform thousands of similarity comparisons per second while maintaining sub-millisecond query times.

For each document chunk, the system computes cosine similarity between the chunk embedding and all skill embeddings in the FAISS index. Through extensive empirical testing across diverse document types, we established an optimal threshold of $\tau = 0.35$ for determining skill relevance. This threshold balances precision and recall, capturing relevant skills while minimizing false positives. Skills exceeding this threshold are

retained as candidates, with their similarity scores preserved for subsequent ranking and filtering operations.

To capture implicit skills that may not be directly stated but are evident from context, we employ SpaCy-based syntactic analysis Honnibal et al. (2020) that identifies verb-object patterns indicative of competencies. For instance, phrases like "coordinate international logistics" or "analyze environmental data" are recognized as implicit references to specific skills even when those exact skill names do not appear in the text. The frequency aggregation stage consolidates skill matches across all document chunks, with sophisticated post-processing algorithms removing near-duplicates and resolving conflicts between similar skills. This consolidation process considers both semantic similarity and hierarchical relationships within the ESCO taxonomy to produce a clean, non-redundant skill list that accurately represents the document's competency landscape.

4.5 Validation Framework

Our validation framework employs a comprehensive multi-tier approach that combines synthetic testing for controlled evaluation with real-world validation for practical applicability assessment. The synthetic validation component utilizes 80 carefully constructed documents, equally divided between those containing explicit skill mentions and those with implicit skill references. These documents incorporate 200 curated ESCO skills spanning technical, managerial, and cross-cutting categories, ensuring comprehensive coverage of the skill taxonomy. Each synthetic document includes ground-truth annotations created by domain experts, enabling precise calculation of precision, recall, and F1 scores for different skills extraction scenarios.

The real-world validation extends the evaluation to practical applications through analysis of 120 policy documents sourced from European institutions, including the European Commission, European Parliament, and various regulatory bodies. These documents represent diverse policy areas including environmental regulation, digital transformation initiatives, and workforce development strategies. Expert annotators with backgrounds in human resources, sustainability, and policy analysis provided detailed annotations for both skill identification and SDG alignment, creating a robust benchmark for system evaluation. The validation set deliberately includes challenging document types such as complex multi-stakeholder policies, technical curriculum vitae with specialized competencies, and course descriptions with varying levels of detail and structure. This diversity ensures that performance metrics reflect real-world applicability across the full spectrum of potential use cases.

4.6 Technical Implementation

The system architecture implements a modular design that facilitates both scalability and maintainability while ensuring optimal performance for production deployment. The backend infrastructure is built on Python 3.9, leveraging Flask and FastAPI frameworks to provide robust API endpoints that handle document uploads, processing requests, and result retrieval. This dual-framework approach allows us to combine

Flask's mature ecosystem for web application features with FastAPI's high-performance asynchronous capabilities for computationally intensive operations.

The frontend interface utilizes Dash Parmer et al. (2017) and Plotly Plotly Technologies Inc. (2015) libraries to create interactive visualizations that enable intuitive exploration of analysis results. This technology stack provides responsive, web-based dashboards that require no client-side installation while supporting complex interactive features such as dynamic filtering, drill-down capabilities, and real-time updates. The natural language processing engine combines multiple specialized tools, with Sentence-Transformers providing state-of-the-art semantic embeddings and spaCy delivering robust syntactic analysis for grammatical pattern recognition. This combination enables the system to capture both semantic similarity and structural patterns indicative of skills.

Vector storage and retrieval operations rely on FAISS indices optimized for different data types. The primary skill index (skill_index.faiss) contains embeddings for the complete ESCO taxonomy, while a separate course index (course_index.faiss) enables rapid matching of identified skills to relevant educational opportunities. Additional indices for occupations and SDG descriptions support comprehensive analysis across multiple dimensions. The caching infrastructure employs Redis for session management, maintaining user state across multiple analysis sessions while enabling rapid retrieval of previous results. An LRU cache for embeddings significantly reduces computational overhead by storing frequently accessed vector representations, particularly beneficial when processing documents with repetitive content.

Data processing operations leverage the scientific Python ecosystem, with Pandas The pandas development team (2020) providing efficient data manipulation capabilities for handling large skill datasets and NumPy Harris et al. (2020) enabling optimized numerical operations for similarity computations and statistical analysis. The web interface supports intuitive interaction through drag-and-drop file uploads with automatic format detection, real-time processing feedback including progress bars and status updates, and comprehensive visualization options. These include pie charts for skill distribution analysis, bar charts for occupation and SDG alignment visualization, and detailed tables for course recommendations with sortable columns and export capabilities. The entire system is containerized using Docker, ensuring consistent deployment across different environments while facilitating horizontal scaling to handle increased demand.

5 Discussion

The results demonstrate the framework's effectiveness in automating complex skill analysis tasks while maintaining high accuracy. On our benchmark, skills extraction performance approaches expert-level benchmarks (overall F1 = 0.963; up to 0.96 on specific categories), validating the combination of semantic embeddings with chunk-based analysis. Particularly noteworthy is the ability to detect implicit skills, which are often overlooked in traditional keyword-based approaches but are crucial for comprehensive workforce assessment.

The strong performance on environmental SDGs (F1 = 0.809) aligns with global priorities for green transition, though lower precision for social SDGs (0.682) indicates room for improvement. This disparity likely reflects the current emphasis on environmental sustainability in policy documents and training data, suggesting the need for expanded datasets covering social dimensions of sustainable development.

From a practical perspective, the framework's rapid processing supports near–real-time policy analysis and workforce planning on typical documents. Organizations can quickly identify skill gaps, align training programs with SDG objectives, and track progress toward sustainability goals. The case study on the EU deforestation policy demonstrates how the framework can extract actionable insights from complex policy documents, identifying not just explicit skill requirements but also implicit competencies needed for successful implementation.

However, several limitations warrant consideration. The current optimization for European contexts may limit global applicability, requiring adaptation for different regional skill frameworks and languages. Additionally, the system's reliance on pre-trained language models introduces potential biases that must be actively monitored and mitigated through regular audits and diverse training data. A key limitation is that evaluation focused on European policy contexts and English-language documents; broader application will require adapting the framework to multilingual and regional skill taxonomies.

6 Conclusion

This research presents a significant advancement in applying artificial intelligence to sustainable workforce development. By automatically extracting, analyzing, and aligning skills with SDG targets, our framework addresses a critical bottleneck in achieving global sustainability goals. The high accuracy achieved across multiple evaluation metrics validates the technical robustness of the approach, while real-world applications demonstrate practical value for diverse stakeholders.

The framework should be viewed as a decision-support tool rather than a definitive classifier: results are intended to augment, not replace, expert judgment. This human-in-the-loop design ensures actionable insights while maintaining accountability.

The integration of SDG alignment into automated skill analysis represents an important advancement in workforce planning, enabling organizations to make decisions that simultaneously address business needs and sustainability imperatives. As we approach the 2030 deadline for SDG achievement, such tools become increasingly vital for accelerating the transformation of human capital.

The pipeline is designed to process multiple document formats (HTML, PDF, DOCX, XML); however, in its current online implementation the system supports HTML documents only. This limitation does not undermine the framework's broader applicability but highlights the importance of continued development and deployment refinement.

Overall, the system represents a practical, scalable step toward SDG-aligned workforce planning. By combining semantic skills extraction with sustainability analysis,

it offers a replicable model that can evolve through expanded datasets, multilingual support, and community-driven development.

Future developments will focus on extending document format support, enhancing social SDG detection, and incorporating federated learning approaches for privacy-preserving analysis. The planned open-source release of core components will facilitate broader adoption and community-driven improvements, potentially establishing this framework as a standard tool for SDG-aligned workforce development. Its integration into the Global Climate Hub's decision infrastructure demonstrates its readiness to support national and international policy implementation aligned with the UN 2030 Agenda.

Acknowledgements. We thank the AE4RIA network partners for providing access to course databases and policy documents. Special recognition goes to the European Commission's ESCO team for maintaining the comprehensive skills taxonomy that underpins this work.

Declarations

- Funding: This research received funding from the European Union's Horizon 2020 innovation action programme under grant agreements No.101037424 (ARSINOE) and No.101037084 (IMPETUS).
- Conflict of interest/Competing interests: The authors declare no conflict of interest.
- Ethics approval and consent to participate: Not applicable - no human participants or animal subjects were involved.
- Consent for publication: All authors consent to publication.
- Data availability: ESCO taxonomy data is publicly available at https://ec.europa.eu/esco/portal. Synthetic test data available upon request.
- Materials availability: Not applicable.
- Code availability: Core system components will be released as open-source following publication.
- Author contribution: PK conceptualized the overall research framework and provided strategic guidance. CP designed the validation methodology, led the SDG category-level evaluation and interpretation, and critically revised the manuscript for framing and structure. CL supported the system architecture design and ESCO dataset integration. GF led the machine-learning implementation, developed the skills extraction and SDG alignment modules, and engineered the web application and deployment. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

References

United Nations (2015) Transforming our world: The 2030 Agenda for Sustainable Development. United Nations, New York

International Labour Organization (2018) World Employment and Social Outlook 2018: Greening with Jobs. ILO Publications, Geneva

International Labour Organization (2019) Skills for a Greener Future: A Global View. ILO Publications, Geneva

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Association for Computational Linguistics, pp 4171–4186

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, pp 3982–3992

Wang W, Bao H, Dong L, Wei F, Zhou M (2020) MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In: Advances in Neural Information Processing Systems (NeurIPS), pp 5776–5788

Johnson J, Douze M, Jégou H (2019) Billion-Scale Similarity Search with GPUs. IEEE Transactions on Big Data 7(3):535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Koundouri P, Landis C, Toli E, Papanikolaou K, Slamari M, Epicoco G, Hui C, Arnold R, Moccia S (2023) Twin Skills for the Twin Transition: Defining Green and Digital Skills and Jobs. AE4RIA White Paper, ATHENA Research Center, Sustainable Development Unit

Koundouri P, Landis C, Koltsida P, Papadaki L, Toli E (2024) Preparing the Maritime Workforce for the Twin Transition: Skill Priorities and Educational Needs. DEOS Working Papers 2417, Athens University of Economics and Business

European Commission (2021) ESCO – European Skills, Competences, Qualifications and Occupations. URL https://ec.europa.eu/esco/portal/home, accessed: 2024-12-15

Honnibal M, Montani I, Van Landeghem S, Boyd A (2020) spaCy: Industrial-Strength Natural Language Processing in Python. Explosion AI. URL https://spacy.io

Richardson L (2023) Beautiful Soup documentation. URL https://www.crummy.com/software/BeautifulSoup/, accessed: 2025-10-04

Felkner M and contributors (2024) PyPDF2: A Pure-Python PDF Library. URL https://pypdf2.readthedocs.io/, accessed: 2025-10-04

python-docx contributors (2024) python-docx: Create and Update Microsoft Word .docx Files. URL https://python-docx.readthedocs.io/, accessed: 2025-10-04

Plotly Technologies Inc. (2015) Collaborative Data Science with Plotly. URL https://plot.ly, accessed: 2025-10-04

Parmer C, Krishnan S, et al. (2017) Dash: A Web Application Framework for Python. Plotly. URL https://dash.plotly.com, accessed: 2025-10-04

Harris CR, Millman KJ, van der Walt SJ, et al. (2020) Array Programming with NumPy. Nature 585:357–362. https://doi.org/10.1038/s41586-020-2649-2

The pandas development team (2020) pandas-dev/pandas: Pandas. Zenodo. https://doi.org/10.5281/zenodo.3509134

Koundouri P, Aslanidis PS, Dellis K, Plataniotis A, Feretzakis G (2025) Mapping Human Security Strategies to Sustainable Development Goals: A Machine Learning Approach. Discover Sustainability 6:96. https://doi.org/10.1007/s43621-025-00883-w