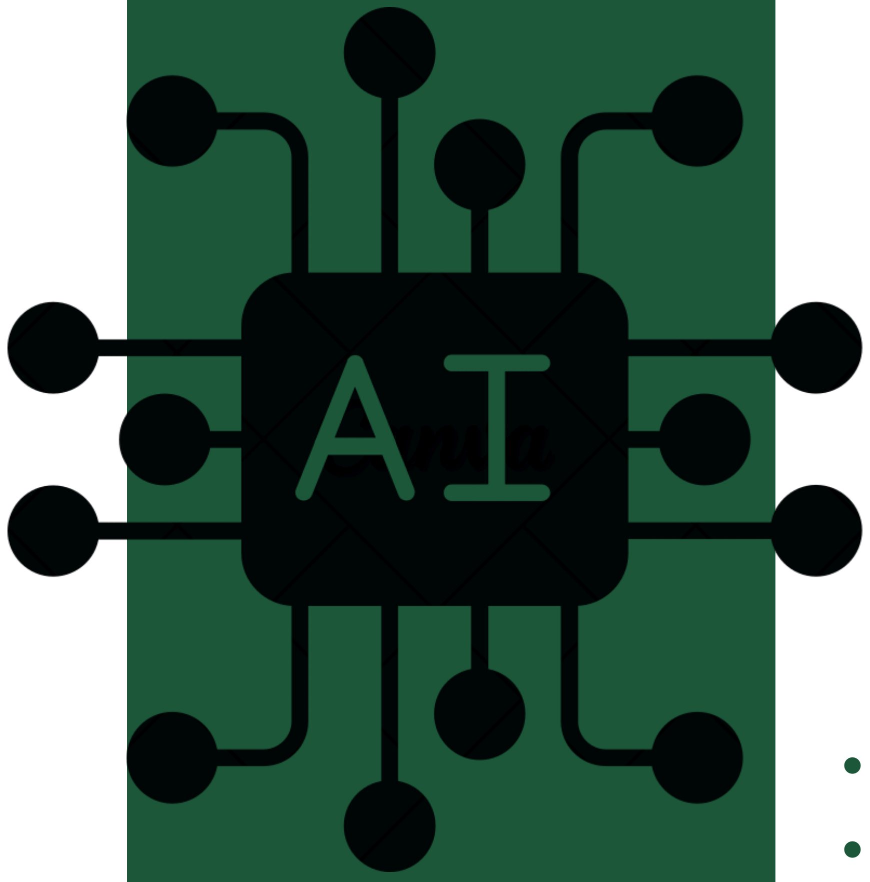
CDI Scotland Student Conference 2025

USING MACHINE LEARNING TO ENHANCE THE PROVISION OF LABOUR MARKET INTELLIGENCE



Agenda

01 Problem Statement

02 Labour Market Analysis

03 Methods

04 Important Findings

05 Current and Future Work

Problem Statement

Considering structural changes as well as ongoing environmental transition, how jobseekers and policymakers can access labour market information efficiently.

Technical Advancements → **Changing Labour Market** → **Skill Shortages**

Brynjolfsson, McAfee (2014) Mezzanzanica, Mercorio (2019)

Research Questions

RQ 1
How has AI been used in similar context, environments and applications (e.g., other countries) for LMI

RQ 2
How data collected from job advertisements can be used to develop explainable ML models for LMI that can enhance decision-making?

RQ 3
How to improve competence extraction from job posting data using ML methods

RQ 4
What are skills and other requirements composition for green jobs?

Current Approaches

Labour Market Surveys

- Costly
- Limited Sample size
- Transparency and Interpretability
- Contextual understanding
- Response Bias
- Takes Time
- Requires pre-requisite knowledge
- Felxibility and adaptability
- Subjective interpretation

Labour Force Survey (ONS)

Employee Skills Survey (UK GOV)

Survey of Adult Skills (OECD)

Automated Approaches

- Explainability
- Large-scale Data Analysis
- Ethical Considerations
- Unbiased Analysis
- Quasi Real-time Insights
- Automated Data collection
- Data Quality

SkillSpan Zhang, Jensen, Sonniks, Plank (2022)

Smith, Weiler, Braschler, (2021)

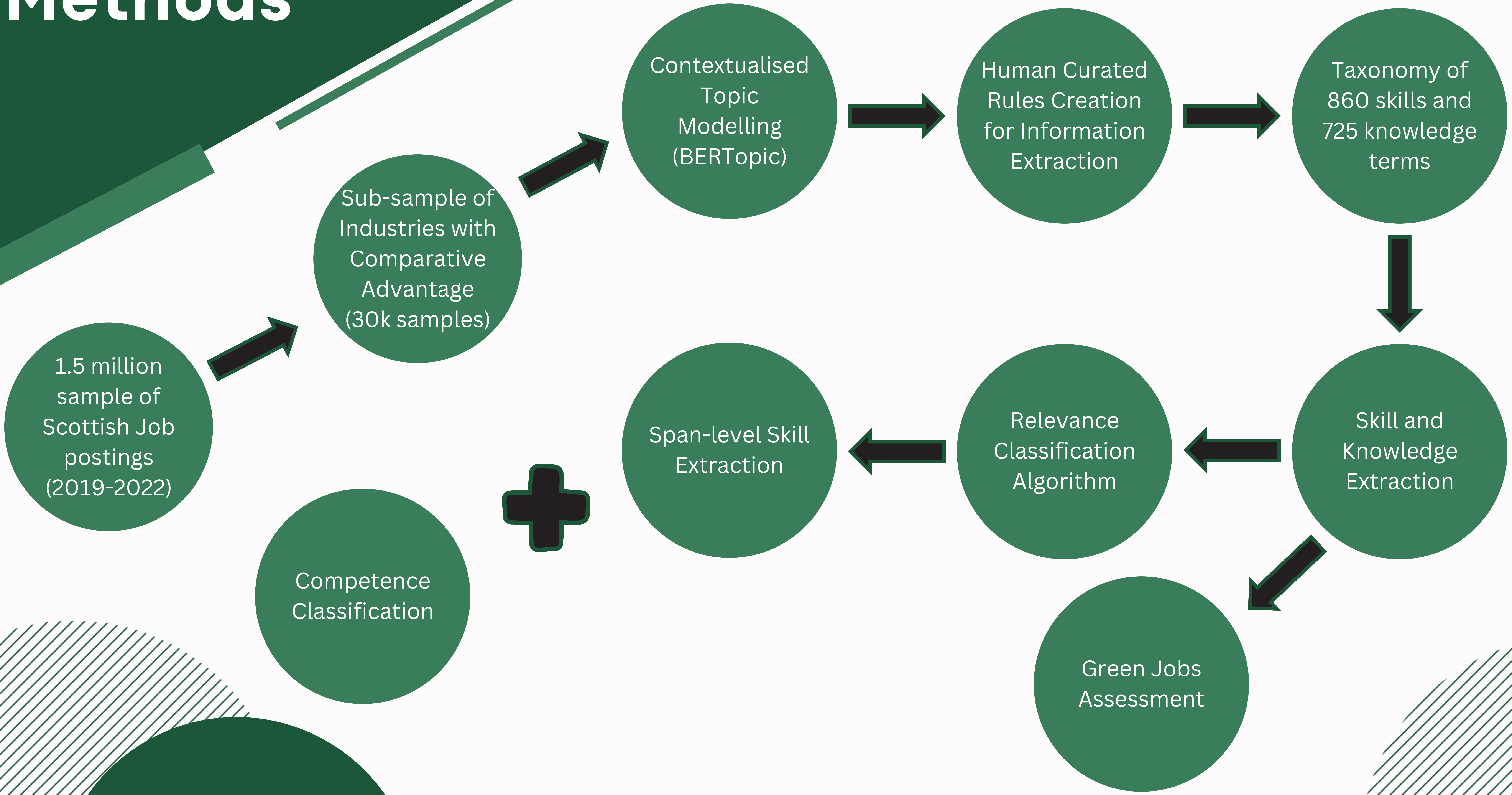
Xu Tong et al., (2018)

Xu, Haoyu et al. (2017)

Burning Glass (2019) (2021)   Fareri, Silvia et al. (2020)

Mezzanzanica, Mercorio (2019). Colombo, Mercorio (2018)

Methods


1.5 million sample of Scottish Job postings (2019-2022)
Sub-sample of Industries with Comparative Advantage (30k samples)
Contextualised Topic Modelling (BERTopic)
Human Curated Rules Creation for Information Extraction
Taxonomy of 860 skills and 725 knowledge terms
Skill and Knowledge Extraction
Relevance Classification Algorithm
Span-level Skill Extraction
Competence Classification
Green Jobs Assessment

Relevance Classification

F1-Scores

Test	Inner-Sample	Cross-Sample	Cross-Sample
Sample	Train: Our Data (80%) Test: Our Data (20%)	Train: Our Data Test: SkillSpan + Sayfullina (20%)	Train: SkillSpan + Sayfullina Test: Our Data (20%)
Irrelevant Class	0.93	0.79 (+7%)	0.72
Relevant Class	0.96	0.83	0.83
Overall	0.94	0.81 (+3%)	0.78

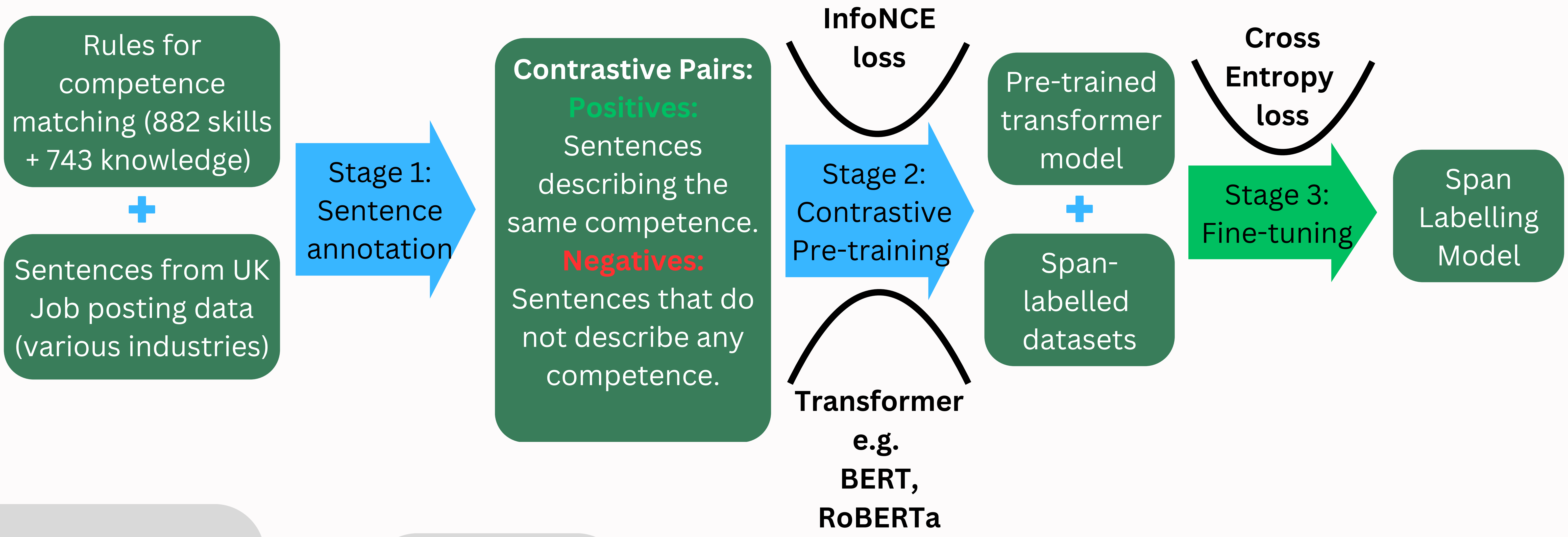
Sentences from Job postings

Proficiency in Python and machine learning packages

Our company hires electrical engineers and is commited to sustainable technologies

Zhang et al., 2022 Sayfullina et al., 2018

Span-level Skill Extraction



A deep understanding of Python programming language.

Vs

Must be able to use Python packages such as PyTorch.

Gunel et al. (2020) Yan et al. (2021) Zhang et al. (2024) Decorte et al. (2023) Oord et al. (2018)

Results

Contrastive Pre-training

Span F1-score Performance Across Datasets

Baselines:
- **Vanilla Fine-tunning**
- **NNOSE (Datastore Method)**

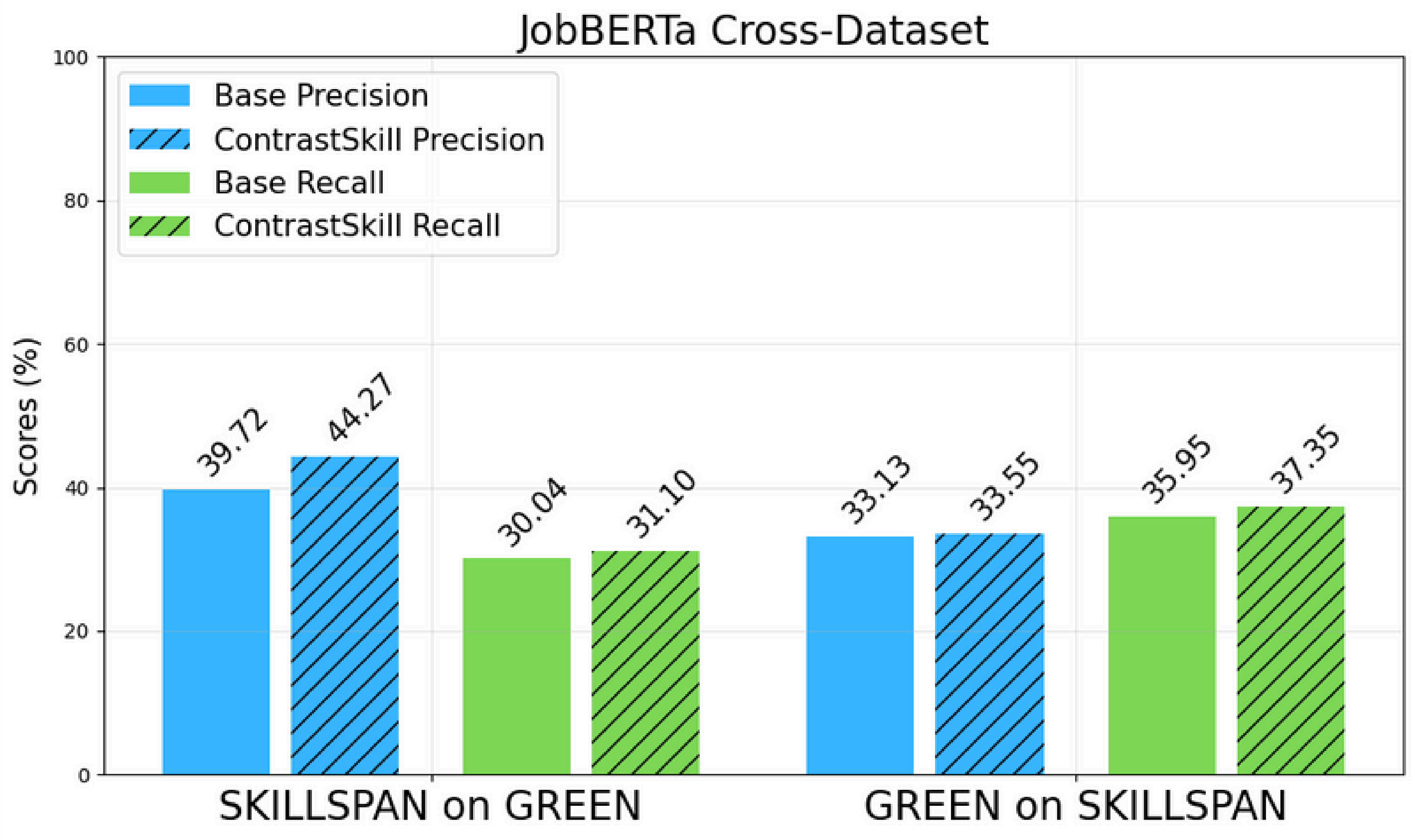
Models:
- **Bert-base-uncased**
- **Roberta-base**
- **JobBERTa (job domain adapted Roberta)**

Datasets:
- **Green (Finance, Healthcare IT)**
- **SkillSpan (Tech)**
- **Sayfullina (vairous industries, focus on soft skills)**

Variant	**Datasets Span-F1 (%)**		
	GREEN	SKILLSPAN	SAYFULLINA
	BERT [13]		
Base	46.40	60.06	93.32
NNOSE [67]	$42.92_{-3.48^*}$	$61.90_{+1.84^*}$	$91.95_{-1.37^*}$
ContrastSkill	$46.64_{+0.24^\dagger}$	$60.81_{+0.75^{*\dagger}}$	$93.17_{-0.15^\dagger}$
	RoBERTa [39]		
Base	43.02	66.15	**93.88**
NNOSE	$44.30_{+1.28^*}$	$65.92_{-0.23}$	$93.36_{-0.52}$
ContrastSkill	$44.68_{+1.66^*}$	$\mathbf{66.43}_{+0.28}$	$93.76_{-0.12}$
	JobBERTa [67]		
Base	51.86	64.12	93.54
NNOSE	$53.50_{+1.64^*}$	$65.96_{+1.84^*}$	$93.12_{-0.42}$
ContrastSkill	$\mathbf{53.96}_{+2.1^*}$	$65.71_{+1.59^*}$	$93.53_{-0.01}$

Results: Cross-Dataset Setting

Contrastive Pre-training

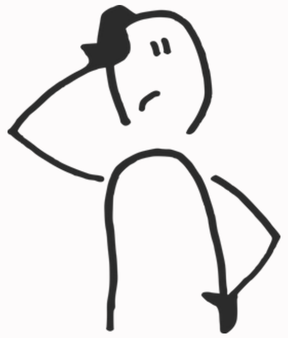


+ 2.32% F1

+ 0.88% F1

COMPETENCE CLASSIFIER

A WORK IN PROGRESS



The Problem

- Scarcity of High-Quality Annotated Data
- Rapidly Emerging Skills
- Industry Specific Data
- Costly Prediction Models

Contrastive Learning

The Solution

ESCO Taxonomy of Skills and Knowledge (Definitions)

+

Examples of Competences from text

=

Example-Definition Alignment

Pros and Cons

- Scalable and Generalisable Competence Classification
- Reduced Reliance on continuous annotation

- Novel Approach (Quality?)
- The importance of diversified examples

Current Work-Labour Market Analysis

01
Generated Rules used to extract skills and knowledge from collected job postings

02
Targeting Green Jobs as defined by the CESAP Pathfinder SDS Report 2023

03
Identification of unique skills and essential competenices in light of broader industry trends

04
Policy Recomendations

Skills Development Scotland (2023)

LinkedIn

Aleksander
Bielinski

THANK YOU

a.bielinski@napier.ac.uk
https://x.com/pecuniafactorem

Summary

Merging ML with Human expertise to ensure efficiency and explainability

Different approaches to skill extraction and classification

Novel ways to take an advantage of different data granularity

Green jobs assessment

REFERENCES

- Zhang, M., Jensen, K.N., Sonniks, S.D. and Plank, B., 2022. SkillSpan: Hard and soft skill extraction from English job postings. arXiv preprint arXiv:2204.12811.
- Sayfullina, L., Malmi, E. and Kannala, J., 2018. Learning representations for soft skill matching. In Analysis of Images, Social Networks and Texts: 7th International Conference, AIST 2018, Moscow, Russia, July 5–7, 2018, Revised Selected Papers 7 (pp. 141-152). Springer International Publishing.
- Skills Development Scotland Pathfinder Report- https://www.skillsdevelopmentscotland.co.uk/media/5hcorb5t/cesap-pathfinder-wp1-report.pdf
- Erik Brynjolfsson and Andrew McAfee. 2014. The second machine age: Work, progress, and prosperity in a time of brilliant technologies. WW Norton & company.
- Mario Mezzanzanica and Fabio Mercorio. 2019. Big data for labour market intelligence: an introductory guide. European Training Foundation.
- Jens-Joris Decorte, Severine Verlinden, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Extreme multi-label skill extraction training using large language models. arXiv preprint arXiv:2307.10778.
- Mike Zhang, Rob van der Goot, Min-Yen Kan, and Barbara Plank. 2024. Nnose: nearest neighbor occupational skill extraction. arXiv preprint arXiv:2401.17092.
- Benjamin Clavié and Guillaume Soulié. 2023. Large language models as batteries-included zero-shot esco skills matchers. arXiv preprint arXiv:2307.03539.
- Fareri, Silvia et al. (2020). “Estimating Industry 4.0 impact on job profiles and skills using text mining”. In: Computers in industry 118, p. 103222.
- Colombo, Emilio, Fabio Mercorio, and Mario Mezzanzanica (2018)
- Smith, E., Weiler, A. and Braschler, M., 2021. Skill extraction for domain-specific text retrieval in a job-matching platform. In Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12 (pp. 116-128). Springer International Publishing.
- Xu, Haoyu et al. (2017). “JCTC: A Large Job posting Corpus for Text Classification”. In: arXiv preprint arXiv:1705.06123.
- Xu, Tong et al. (2018). “Measuring the popularity of job skills in recruitment market: A multi- criteria approach”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. 1.
- Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.
- Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. arXiv preprint arXiv:2011.01403.
- Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: a contrastive framework for self-supervised sentence representation transfer. arXiv preprint arXiv:2105.11741
- Amirhossein Herandi, Yitao Li, Zhanlin Liu, Ximin Hu, and Xiao Cai. 2024.Skill-llm: repurposing general-purpose llms for skill extraction. arXiv preprint arXiv:2410.12052.
- Burning Glass (2019). Mapping the Genome of Jobs: The Burning Glass skills taxonomy.
- Papoutsoglou, Maria, Apostolos Ampatzoglou, et al. (2019). “Extracting knowledge from on- line sources for software engineering labor market: A mapping study”. In: IEEE Access 7, pp. 157595–157613.