Universiteit Leiden

Master Computer Science

Leveraging AI for the Development of a Standardized Multilingual Education Taxonomy Aligned with Job Market Demands

Name:	Abed Alrahman Hettini
Student ID:	s3603970
Date:	23/04/2025
Specialisation:	Artificial Intelligence
1st supervisor:	Niels van Weeren
2nd supervisor:	Suzan Verberne

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgements

First, I would like to express my deepest gratitude to my supervisor, Niels van Weeren, for his exceptional guidance throughout this research project. His unwavering support enabled me to combine my master's thesis with a rewarding internship at Randstad, where I gained invaluable experience through real-world experiments and interactions with inspiring professionals. This opportunity profoundly enriched my research, and I am truly thankful for this transformative academic and professional experience.

I am equally grateful to my supervisor, Suzan Verberne, for her valuable feedback and academic supervision during the writing process. Her insightful comments and collaborative approach were instrumental in shaping this thesis.

My sincere thanks go to the Randstad team for their warm welcome and continuous support during my internship. A special acknowledgement goes to Cecile Ramombordes for fostering both my professional growth and thesis development.

Last but not least, to my family – my Parents, Kholoud, Amir, Simaza – and my friend Yazan, your emotional support and patience during this journey were my foundation. This accomplishment belongs as much to you as it does to me.

Abstract

The lack of a standardized education taxonomy presents a significant challenge in talent management, particularly for global companies like Randstad, which operates across multiple countries with distinct education systems. To address this issue, this research explores the use of transformer-based NLP models for education taxonomy generation and matching educational qualifications with relevant skill taxonomies.

We first develop a multilingual, structured education taxonomy by fine-tuning transformer models: BERT, ModernBERT, mBERT, and XLM-R. Our results indicate that mBERT achieves the highest classification accuracy, demonstrating its ability to generalize across different education systems. Next, we investigate how education taxonomies can be aligned with structured skill taxonomies using two NLP-based approaches: (1) NER-based skill extraction and (2) embedding-based matching. Our experiments show that embedding-based approaches yield better alignment with occupational skills, offering a scalable solution for skill-based job matching.

Furthermore, we evaluate the model's performance on real-world education data from Belgium and Italy, comparing BERT-predicted taxonomies with human-assigned classifications. While Belgium exhibits moderate alignment, the results for Italy reveal discrepancies, where the model's predictions often outperform human classifications—highlighting the potential limitations of expert labeling.

Overall, our findings demonstrate that transformer models can successfully generate a standardized education taxonomy that generalizes across countries and institutions, ensuring cross-border comparability. Additionally, NLP-driven approaches can effectively bridge the gap between education and skill taxonomies, enhancing automated job matching.

Contents

1 Introduction

The landscape of education is continually evolving, driven by technological advancements, shifting societal needs, and the dynamic demands of the global job market. Furthermore, education systems worldwide have developed independently, leading to a diverse array of qualifications, standards, and terminologies. This diversity has historically posed challenges for comparability and standardization [51]. With globalization, the need for a unified framework to categorize educational qualifications has become increasingly apparent, and developing comprehensive education taxonomies that classify and organize educational qualifications is essential for creating a cohesive and navigable educational ecosystem.

1.1 Problem statement

Educational programs across various countries are provided by many institutions such as universities, high schools, vocational schools, and specialized training centers. Each of these institutions classifies and categorizes their programs based on national, regional, and institutional criteria, often resulting in significant differences in classification methods and standards. These classifications vary not only in terms of educational level but also in program structure, curriculum content, duration, and assessment criteria.

Such variations lead to challenges in comparing educational programs across borders, as stakeholders face difficulties in understanding equivalencies and recognizing qualifications from one country to another. For students and professionals looking to further their education or pursue employment opportunities internationally, the lack of standardized categorization limits their ability to navigate and evaluate program options effectively [10]. Similarly, employers, educators, and policymakers struggle to assess the compatibility of qualifications and skills obtained in different educational systems with the requirements of local labor markets [14].

Randstad, the world's largest HR company, has assisted over 2 million individuals in finding suitable jobs and provided guidance to more than 230,000 clients on talent management [41]. To further enhance its services,

the company plans to create a standardized education taxonomy that will play a pivotal role in classifying educational data in talent profiles, such as resumes, and guiding individuals toward relevant education. This system will ensure that educational qualifications are accurately represented and aligned with job opportunities, regardless of regional variations.

Currently, Randstad faces challenges because there is no unified education taxonomy across its global branches. Countries like the Netherlands, Italy, Sweden, and Belgium each have their own distinct education systems and taxonomies, which complicates the process of managing and comparing qualifications consistently. Therefore, there is a critical need to develop a standardized taxonomy that accommodates regional differences while being accessible in the local languages of each country.

The company intends to integrate this standardized education taxonomy into its global core applications to improve job matching accuracy and enhance talent profiles. For consultants, the taxonomy enables precise specification of educational requirements in job orders. Consultants can set desired (minimum) qualification levels, fields of study, and specializations, with options to mark certain qualifications. This flexibility allows for a more refined approach to role requirements. Additionally, consultants can filter educational data to display only relevant qualifications, such as showing only specializations within a selected field like "Economics" streamlining the selection process.

For job-seekers using company websites [41], the taxonomy supports more detailed profile creation by allowing individuals to input their educational credentials in a standardized format. Job-seekers can specify their degree completion status and the institutions where their qualifications were obtained, resulting in a more comprehensive and accurate academic profile. This integration of a standardized education taxonomy would enhances Randstad's platforms by fostering consistency, minimizing ambiguity, and improving the alignment of educational qualifications with job opportunities, ultimately optimizing Randstad's job-seekers matching and workforce solutions [11].

Therefore, there is a problem lies in the need for a standardized education taxonomy that not only aligns and categorizes programs according to comparable levels and criteria but also supports seamless data integration and retrieval. Such a taxonomy would streamline the identification and

comparison of educational programs across countries, ultimately enabling individuals and institutions to make informed decisions regarding education and workforce alignment on a global scale.

1.2 Research approach

Our research is carried out with an internship at Randstad company. It leverages a company use case requiring the development of an education taxonomy for official educational program data from the Netherlands, Belgium, and Italy, with potential future expansion to additional countries, such as Sweden, France, the United States, Portugal, and Switzerland.

The research further aims to establish a method for aligning the generated education taxonomy with existing skills taxonomies, such as the Open Skills Framework by Lightcast [30]. This alignment represents the academic contribution of the research and is intended to support future Randstad applications by improving the interoperability between educational and skills data.

1.3 Research questions

To address the challenges associated with developing a standardized education taxonomy using techniques that provide scalable solutions for analyzing unstructured education data, automating pattern recognition, and resolving semantic ambiguities like NLP and machine learning techniques, this study poses several research questions. These questions aim to investigate the effectiveness of NLP models, the adaptability of the taxonomy across multiple languages and educational systems, and the potential alignment with skills taxonomies. Each question is designed to guide the research toward practical solutions that enhance taxonomy accuracy, interoperability, and applicability in real-world settings.

The **main** question that we're proposing to address in this research is:

"***What is the effectiveness of transformers in creating a standardized education taxonomy that is both multilingual and adaptable***

to diverse job market demands across different countries?"

To achieve this, the research addresses several key supporting questions:

In the context of speed and quality, **[RQ1]** *How can transformers improve the efficiency, accuracy, and multilingual adaptability of education taxonomy generation compared to traditional methods?* This question examines how transformer-based NLP models can streamline and enhance the process of developing education taxonomy. It aims to determine whether these models can generate taxonomies more efficiently and accurately than manual or traditional methods, such as:

- Expert-Driven Classification, where education experts manually classify programs into taxonomies based on predefined guidelines (Cedefop, 2017) [9].
- Rule-Based Approaches, which rely on keyword matching and predefined heuristics to categorize programs (Dahler-Larsen, 2018) [16].

Additionally, this question explores whether transformers can effectively address the diverse requirements of education taxonomy, including processing heterogeneous data inputs, handling multilingual education data, and ensuring semantic consistency across different classification levels.

In terms of generalization and standardization, **[RQ2]** *How can an education taxonomy be standardized and generalized across different countries and institutions to ensure cross-border comparability?* This question explores both generalization and standardization in the development of education taxonomy to ensure cross-border comparability across different countries and institutions. It examines how an education taxonomy model, initially designed for one country's system, can be adapted and extended to accommodate the educational structures of other countries. Since education systems vary significantly in their qualification frameworks, levels, and program structures, it is essential to develop a flexible and scalable taxonomy that can be mapped across multiple national contexts.

Focusing on the job market alignment, **[RQ3]** *How closely are the education and skills taxonomies related and how can we establish a link between them*

using NLP techniques? This question examines the connection between educational and skills taxonomies and explores how NLP can create meaningful associations between them, ensuring that educational qualifications align with skill requirements in the job market.

1.4 Contributions

1.4.1 Academic contributions

1. Implementation of a model that can generate a standardized education taxonomy that can be adapted to multiple countries and languages, addressing the complexity of diverse educational systems.
2. Investigation of how NLP enhances the speed and quality of taxonomy generation compared to manual classification.
3. Evaluation of transformer models like BERT, multilingual BERT, ModernBERT and XLM-R on a multilingual real-data.
4. Creation of benchmark datasets for taxonomy validation to evaluate and validate taxonomy generation and alignment methods.
5. Implementation of a methodology to align educational taxonomies with existing skills taxonomies libraries using NLP techniques.

1.4.2 Business contributions

1. Integration of a taxonomy system tailored to Randstad's needs, facilitating efficient classification, organization, and retrieval of educational data across multiple countries.
2. Application of the generated taxonomies in Randstad systems (real-world systems), providing valuable feedback from expert users to evaluate and refine the effectiveness and accuracy of the taxonomies.
3. Automating manual classification processes will save Randstad significant time while providing more accurate results, thus improving

the efficiency and effectiveness of the company's digital platforms and applications.

4. Aligning educational qualifications with market-relevant skills, helping Randstad offer better job matching and career planning services.

2 Background & Related work

2.1 Education Systems

Education systems are the foundational structures through which societies organize and provide education to individuals at various stages of their lives. These systems vary significantly between countries and regions and are shaped by cultural, historical, economic, and political factors. Despite their differences, most education systems share common objectives: impart knowledge, foster critical thinking, develop skills, and prepare individuals for participation in the workforce and society [50].

Typically, education systems are organized into several levels, which may include [50]:

- Early childhood education, this foundational stage focuses on the cognitive, social, and emotional development of young children, usually aged 3 to 6. It may include preschool or kindergarten programs.
- Primary or elementary education is the first stage of formal schooling, typically spanning ages 6 to 12. The curriculum emphasizes literacy, numeracy, basic science, and social studies.
- Secondary education is divided into lower and upper secondary stages. Provides general, vocational, or technical education to students aged 12 to 18, preparing them for higher education or employment.
- Higher education includes universities, colleges, and vocational training centers that offer undergraduate, graduate, and professional education. Tertiary education aims to deepen knowledge and skills in specific fields, promoting research, innovation, and workforce readiness [47].
- Lifelong learning and adult education, recognizing the need for continuous skill development, many systems offer adult education programs, certifications, and lifelong learning opportunities to adapt to changing career demands [46].

Although the basic structure of education systems is similar, their implementation differs significantly between countries [36]. One major area of variation are the curricula and standards, which are set by national or regional authorities. These curricula can differ in content, depth, and focus, with some systems emphasizing science, technology, engineering, and mathematics, while others prioritize humanities and social sciences. Another key difference lies in educational pathways, where systems can offer multiple pathways, such as academic, vocational, or technical options. This allows students to pursue an education that aligns with their interests and career aspirations, providing flexibility within the overall structure. Qualification frameworks also vary, as many countries implement these frameworks to classify educational levels and align them with skill requirements. A notable example is the European Qualifications Framework (EQF) [13], which harmonizes qualifications across European nations, facilitating better comparability and mobility. Finally, the governance and funding of educational systems can differ greatly. Some systems are managed by local or regional authorities. Funding models range from primarily publicly funded systems to those that depend significantly on private institutions, reflecting the diverse approaches to financing education across different countries.

Education systems face several key challenges that impact their effectiveness and inclusivity. The balance between standardization and flexibility remains significant issues, while standardization facilitates comparisons and mobility across different regions and countries, flexibility is necessary to adapt education to local needs and contexts, creating a tension that education systems must manage carefully [52]. Skill alignment is another critical challenge, as many educational systems struggle to ensure that the skills and qualifications they provide meet the evolving needs of the labor market [23]. This misalignment can lead to skill gaps and unemployment, affecting both individual career prospects and broader economic health. Lastly, multilingual and multicultural integration presents a complex challenge, especially in societies with diverse linguistic and cultural demographics [24]. Education systems must accommodate this diversity while promoting social cohesion, which requires thoughtful policies and adaptable curricula.

A Qualifications Framework (QF) is a structured system that classifies and describes qualifications based on learning outcomes, competencies, and levels of proficiency. These frameworks provide a basis for recognizing and

validating skills, ensuring transparency in education and labor markets [10]. In the context of this research, a standardized education taxonomy aligns closely with the objectives of qualifications frameworks. Since Randstad aims to integrate a unified taxonomy for classifying educational data across multiple countries, leveraging principles from existing QFs can enhance cross-border comparability. The following subsections will explore various national and international qualifications frameworks, highlighting their structures, differences, and implications for our taxonomy-based classification approach.

2.2 European Qualifications Framework (EQF)

The European Qualifications Framework (EQF) is a tool designed to promote the comparability of qualifications in European countries, facilitating mobility for learners and workers within the European Union [13]. Introduced in 2008, the EQF serves as a common reference framework that links different countries' national qualifications frameworks (NQFs), allowing qualifications to be translated into a common European language. This helps employers and educational institutions across the EU understand the levels of different qualifications.

The EQF is structured into eight levels, each defined by a set of descriptors indicating the learning outcomes, such as knowledge, skills, and competencies, that a learner must achieve to obtain a qualification at that level [13]. These levels range from basic (Level 1) to advanced qualifications (Level 8), which correspond to doctoral degrees. The descriptors are designed to be neutral in the learning process, focusing on what the learner knows, understands, and can do at the end of a learning process, rather than how or where the learning took place [9].

The implementation of the EQF across Europe has involved aligning national qualification frameworks (NQFs) with the EQF levels. This alignment allows for a more transparent comparison of qualifications and supports the recognition of qualifications across borders. By 2020, most EU countries had completed or were near completion of the reference of their national frameworks to the EQF [19].

The EQF has significantly impacted lifelong learning policies by encourag-

ing the validation of non-formal and informal learning [9]. This validation process ensures that skills acquired outside traditional education systems, such as work experience, are recognized and can contribute to formal qualifications. This is crucial to improve the employability of individuals, especially in a dynamic labor market [7].

The Netherlands Qualifications Framework (NLQF) is an example of how a national framework aligns with the EQF. The NLQF categorizes all qualifications within the Netherlands into eight levels, which correspond directly to the EQF levels. This alignment ensures that Dutch qualifications are transparent and comparable throughout Europe, facilitating mobility for learners and workers (NCP NLQF, 2019) [35]. NLQF levels are defined by learning outcomes in terms of knowledge, skills, and competencies. These levels range from entry-level qualifications to advanced degrees, including doctoral levels, making it easier to understand the progression and equivalence of qualifications within and outside the Netherlands. Figure 2.1 illustrates the eight levels of NLQF, showing their equivalence to the EQF levels and how various Dutch qualifications, such as secondary education, vocational training, and higher education degrees, are mapped to these levels.

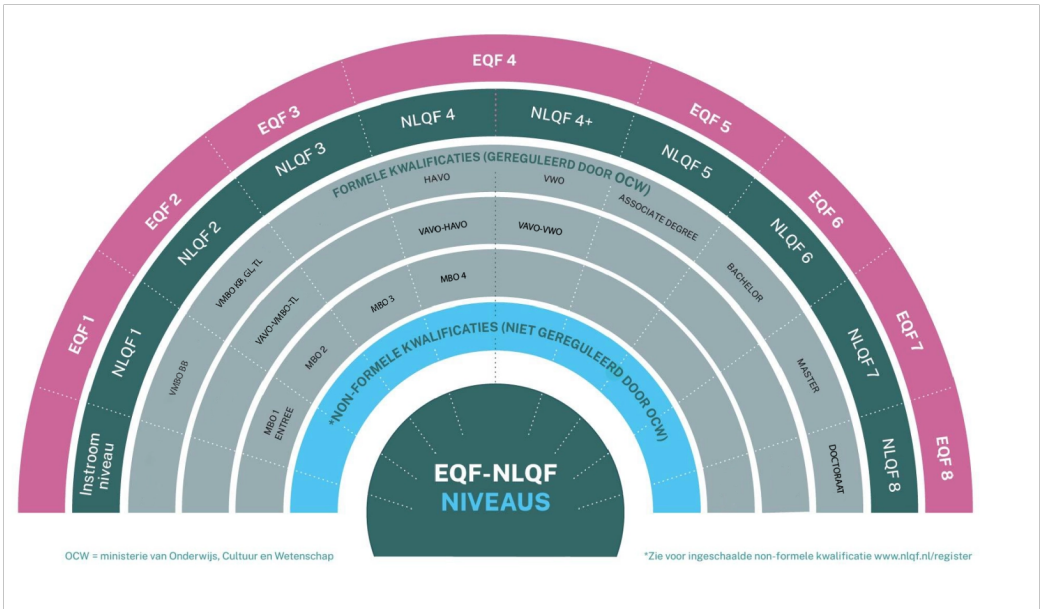

Figure 2.1: Alignment of The Netherlands Qualifications Framework (NLQFs) with the EQF. Source: [35].

2.3 International Standard Classification of Education (ISCED)

The International Standard Classification of Education (ISCED) is a globally recognized framework designed to collect, compile, and analyze education statistics across different countries. As a key member of the United Nations International Family of Economic and Social Classifications, ISCED serves as a reference for organizing education programs and related qualifications by levels and fields of education. Originally developed in the 1970s by the United Nations Educational, Scientific and Cultural Organization (UNESCO), ISCED has undergone several revisions, notably in 2011 and 2013, to better reflect the evolving landscape of global education systems [50]. The 2011 revision focused on refining the classification of education levels, leading to the creation of ISCED-P (program levels) and ISCED-A (attainment levels). These changes enhanced the framework's ability to capture the various stages and achievements within formal education systems. The 2013 revision, known as ISCED Fields of Education and Training (ISCED-F 2013) [49], introduced

a comprehensive classification of education fields, aimed at improving the analysis of education and training by specific areas of study.

ISCED-F 2013 classifies educational programs into detailed fields based on their subject content, providing clear guidelines on what constitutes each field and offering examples to distinguish similar areas. This classification supports better differentiation and clarity in education statistics, helping policymakers and researchers to analyze trends in specific fields of education. The ISCED revisions are products of international collaboration and are formally adopted by the UNESCO General Conference, ensuring their relevance and applicability across member states.

The ISCED offers a detailed and hierarchical framework for classifying educational content into three levels: broad fields, narrow fields, and detailed fields, as illustrated in Figure 2.2.

Broad Fields, these are the highest level of classification in ISCED-F 2013, including 11 categories classified in a general manner.

Narrow Fields, each broad field is divided into several narrow fields, which represent more specific areas within the broad field. These narrow fields bring together closely related disciplines for more focused analysis.

Detailed Fields, the most granular level of classification is the detailed fields, which fall under the narrow fields. They provide the most specific classification of educational content.

Broad field	Narrow field	Detailed field
00 Generic programmes and qualifications	000 Generic programmes and qualifications not further defined	0000 Generic programmes and qualifications not further defined
	001 Basic programmes and qualifications	0011 Basic programmes and qualifications
	002 Literacy and numeracy	0021 Literacy and numeracy
	003 Personal skills and development	0031 Personal skills and development
	009 Generic programmes and qualifications not elsewhere classified	0099 Generic programmes and qualifications not elsewhere classified
01 Education	011 Education	0110 Education not further defined 0111 Education science 0112 Training for pre-school teachers 0113 Teacher training without subject specialisation 0114 Teacher training with subject specialisation 0119 Education not elsewhere classified
	018 Inter-disciplinary programmes and qualifications involving education	0188 Inter-disciplinary programmes and qualifications involving education
02 Arts and humanities	020 Arts and humanities not further defined	0200 Arts and humanities not further defined
	021 Arts	0210 Arts not further defined 0211 Audio-visual techniques and media production 0212 Fashion, interior and industrial design 0213 Fine arts 0214 Handicrafts 0215 Music and performing arts 0219 Arts not elsewhere classified
	022 Humanities (except languages)	0220 Humanities (except languages) not further defined 0221 Religion and theology 0222 History and archaeology 0223 Philosophy and ethics 0229 Humanities (except languages) not elsewhere classified
	023 Languages	0230 Languages not further defined 0231 Language acquisition 0232 Literature and linguistics 0239 Languages not elsewhere classified
	028 Inter-disciplinary programmes and qualifications involving arts and humanities	0288 Inter-disciplinary programmes and qualifications involving arts and humanities
	029 Arts and humanities not elsewhere classified	0299 Arts and humanities not elsewhere classified

Figure 2.2: Hierarchical Structure of ISCED-F 2013, Showing Broad, Narrow, and Detailed Fields of Education. Source: [49].

The ISCED-F framework uses a coding system to identify and organize these fields hierarchically. Each level is assigned a unique numeric code that reflects its position within the hierarchy. Broad fields are assigned two-digit codes (e.g.

01 for Education). Narrow fields are identified by a three-digit code, where the first two digits correspond to the broad field, followed by an additional digit (e.g. 011 for "Education science"). Detailed fields are represented by a four-digit code, where the first three digits align with the narrow field, and the fourth digit provides further specificity (e.g. 0111 for "Curriculum studies").

2.4 Skills taxonomy

A skills taxonomy is a structured framework that organizes and categorizes various skills, competencies, and abilities required across different industries and job roles. It serves as a critical tool for understanding labor market demands, facilitating workforce planning, and guiding education and training programs to ensure alignment with economic needs. Skills taxonomies are designed to capture the dynamic nature of the job market by categorizing skills into hierarchical structures, typically ranging from broad skill categories to more specific skills.

Integrating a skills taxonomy with an education taxonomy involves mapping educational programs and qualifications to the relevant skills they impart. This alignment ensures that educational outcomes are directly linked to labor market requirements, enhancing the relevance and employability of graduates [12].

The Lightcast Skills Taxonomy is a dynamic and comprehensive framework that categorizes over 32,000 distinct skills extracted from job postings, resumes, and professional profiles. Continuously updated to reflect the evolving nature of work, it captures both emerging skills and declining demands. The taxonomy is available in multiple languages and is structured hierarchically, organizing skills into broad categories and narrower subcategories, and also organized into three primary classifications [30]:

- **Common skills:** These include widely applicable skills across various industries and occupations, such as soft skills ("Communication") or general technical skills ("Microsoft Excel"). They encompass both personal attributes and learned competencies.
- **Specialized skills:** These are industry specific skills or those necessary

for performing particular tasks, such as "NumPy" for data analysis or "Hotel Management" in the hospitality sector. Known as technical or hard skills, they cater to specific job functions.

- **Certifications:** These are industry recognized qualifications or standards, such as a "Cosmetology License" or "Certified Cytotechnologist." Certifications serve as formal validation of an individual's expertise in a specific field.

In this thesis, we use the Lightcast Open Skills library to integrate a comprehensive skills taxonomy with an education taxonomy. We choose this library due to its extensive coverage of skills, continuous updates reflecting emerging trends, and its hierarchical structure. This makes it a robust and adaptable tool for accurately mapping educational qualifications to the dynamic demands of the labor market.

2.5 Transformers

Transformer-based language models have revolutionized NLP by enabling contextualized word representations that capture semantic relationships more effectively than traditional word embeddings. The transformer architecture replaces recurrent and convolutional structures with a self-attention mechanism, allowing models to process entire input sequences in parallel while maintaining long-range dependencies.

One of the key innovations in Transformer models is the use of embedding layers, which map input tokens into dense vector representations before passing them through multiple self-attention layers. Unlike static embeddings such as Word2Vec or GloVe, Transformers generate contextualized embeddings, meaning the representation of a word dynamically adapts based on its surrounding context [17].

Embedding-based approaches have demonstrated remarkable effectiveness in various NLP tasks, including text classification, entity recognition, and taxonomy generation [38]. In the context of our research, these transformer embeddings play a crucial role in both education taxonomy generation and

mapping educational qualifications to relevant skills. By fine-tuning pre-trained transformer models, we leverage their rich linguistic representations to enhance classification accuracy and improve cross-border comparability in education systems.

One widely adopted transformer model is BERT (Bidirectional Encoder Representations from Transformers), which plays a crucial role in learning contextualized embeddings that enhance the performance of downstream classification tasks. BERT and its variants form the foundation for many of the models used in this thesis and will be introduced in more detail in next sections.

2.6 Cross-lingual models

The growing global demand for language-agnostic natural language processing NLP systems has led to significant advancements in cross-lingual and multilingual transformer models. These models are designed to support multiple languages simultaneously, allowing for knowledge sharing across linguistic boundaries without the need for parallel data or extensive translation resources. This is particularly relevant in real-world applications such as machine translation, cross-lingual information retrieval, and international classification tasks, including the harmonization of educational taxonomies across countries.

Unlike monolingual models, multilingual NLP models are pretrained on massive corpora spanning dozens or even hundreds of languages. A widely used model in this space is Multilingual BERT (mBERT), which extends the original BERT architecture by jointly training on Wikipedia data in 104 languages [17]. Another state-of-the-art model is XLM-RoBERTa (XLM-R), which outperforms earlier multilingual models through deeper architecture and a more extensive multilingual pretraining corpus [15]. These models learn shared semantic representations that are aligned across languages, making it possible to apply the same model architecture and weights to multilingual tasks without explicit translation.

In this thesis, such cross-lingual capabilities are critical for addressing the challenge of standardizing educational taxonomies across countries. Education

systems in countries like Italy, Belgium, and the Netherlands use different structures, terminology, and classifications. By leveraging multilingual transformers, we aim to bridge these semantic gaps and build classification models that are not only accurate in a single language but generalize well across languages and national contexts.

Additionally, the use of multilingual models supports transfer learning, wherein a model trained on one language (or country's data) can be applied or fine-tuned on another, thereby reducing the data annotation burden. This capability is particularly useful in domains such as education, where labeled datasets are scarce, inconsistent, or manually curated across countries.

These models also contribute to semantic alignment across educational and skills taxonomies, which is essential for facilitating cross-border recognition of qualifications and enabling interoperability in talent platforms, such as the systems used by Randstad. In this regard, multilingual NLP models are foundational to the core objectives of this research, providing the infrastructure needed to map diverse educational qualifications onto a unified framework that is both scalable and internationally applicable.

2.7 Named Entity Recognition (NER)

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that involves identifying and classifying entities within a text into predefined categories such as persons, organizations, locations, dates, and numerical values [34]. It is widely used in various applications, including information extraction, question answering, and knowledge graph construction.

NER plays a crucial role in educational data processing, as it helps in structuring unstructured text by extracting relevant entities such as degree titles, institutions, skills, and occupations. In the context of this research, NER is leveraged to identify and categorize educational qualifications and their associated skills, which facilitates alignment with education taxonomies and improves classification accuracy.

Traditional NER approaches relied on handcrafted rules and dictionar-

ies [42]. However, recent advances in deep learning and transformer-based models [17] have significantly improved NER performance, enabling more accurate and context-aware entity recognition. Pretrained transformer models such as BERT and DeBERTa have been successfully applied to NER tasks, demonstrating state-of-the-art performance across multiple languages and domains.

In this research, NER contributes to the extraction of skills and qualifications from job descriptions and resumes, aiding in the development of a standardized education taxonomy. By applying transformer-based NER models, we enhance the accuracy of educational classification and improve the linkage between education and skills taxonomies.

2.8 Related work

The generation and integration of educational taxonomies with skills taxonomies using Natural Language Processing (NLP) is a novel research area with limited direct studies. However, several adjacent fields provide a foundation for understanding the intersection of educational content classification, taxonomy development, and skills mapping. This section reviews key contributions in related domains, highlighting the gap that this thesis aims to address.

2.8.1 Education taxonomy generation

Educational taxonomies, such as the International Standard Classification of Education (ISCED) [49], have been traditionally developed through manual processes involving educational experts. Recent advances in NLP have opened up opportunities for automating the generation of these taxonomies. In this thesis, ISCED serves as the baseline for the taxonomy framework. We build upon this existing structure by incorporating Natural Language Processing (NLP) techniques to automate the classification of educational programs according to the ISCED taxonomy. Traditional manual classification methods, while effective, are often resource-intensive and subject to human bias, which can limit scalability and efficiency. By leveraging NLP, we aim to streamline

the process of categorizing educational programs, enabling a more efficient and consistent mapping to the ISCED taxonomy. This approach will allow for the automatic classification of large datasets of educational programs, enhancing both the speed and accuracy of taxonomy assignment.

Moreover, the use of NLP opens up possibilities for refining the classification process by identifying subtle patterns and relationships within educational data that might be missed by manual classification. This thesis thus seeks to modernize and enhance the application of the ISCED taxonomy through the integration of advanced NLP methods, bridging the gap between traditional manual classification and emerging technological solutions.

2.8.2 BERT for text classification

The BERT (Bidirectional Encoder Representations from Transformers) model, introduced by Devlin et al. (2019) [17], represents a significant advancement in the field of Natural Language Processing (NLP). BERT's architecture is based on a deep bidirectional transformer, which allows the model to consider the context from both the left and right sides of a word simultaneously, enabling a deeper understanding of language differences. This bidirectional approach contrasts with earlier models that processed text unidirectionally, thus limiting their ability to capture the full context.

BERT is pre-trained on large datasets using a masked language model (MLM) objective, where random words in a sentence are masked, and the model learns to predict these masked words based on the surrounding context. This pre-training phase equips BERT with a robust understanding of language, which can be fine-tuned for various downstream NLP tasks, including text classification.

In text classification tasks, BERT has proven to be highly effective due to its ability to understand complex linguistic structures and long-range dependencies within text. When fine-tuned for classification, BERT uses its pre-trained knowledge to classify text into categories by adding a simple classification layer on top of the pre-trained model. This approach has set new benchmarks for various classification tasks, including sentiment analysis, topic classification, and question answering.

BERT's bidirectional nature allows for more accurate text classification by fully understanding the context in which words appear. The pre-training on large datasets enables BERT to be fine-tuned on smaller datasets for specific tasks, making it versatile and efficient for a wide range of classification problems. State-of-the-Art Performance: BERT has achieved state-of-the-art results on several NLP benchmarks, demonstrating its superior performance in text classification tasks.

In this thesis, using BERT for text classification can significantly enhance the accuracy and reliability of classifying educational content into taxonomies. Its capability to understand nuanced language contexts makes it an ideal tool for handling complex educational texts, ensuring precise taxonomy generation and integration.

2.8.3 Cross-lingual taxonomy

One of the central challenges addressed in this thesis is aligning educational taxonomies across multilingual contexts. Recent research in cross-lingual taxonomy alignment offers valuable insights into this issue.

Zhou et al. (2020) [53] propose a method using bilingual knowledge graph embeddings to align taxonomies across different languages. Their approach leverages the structural and semantic relationships encoded in multilingual knowledge graphs to identify correspondences between taxonomy nodes. This method demonstrates strong potential for use in cross-border educational taxonomy standardization, where similar qualifications may be labeled differently depending on language and national frameworks. The success of embedding-based strategies in this context supports our findings that transformer-based models can generalize across diverse taxonomies by capturing deep semantic similarities.

Similarly, Jiménez-Ruiz et al. (2018) [27] introduce a machine learning framework for multilingual and cross-lingual ontology matching. Their system integrates lexical, structural, and external background knowledge to improve alignment across ontologies in different languages. This research highlights the feasibility of using machine learning—notably NLP-driven—techniques to resolve inconsistencies in semantic representation across systems. This aligns

well with our second research question regarding standardizing education taxonomies across countries and confirms the potential of NLP models to unify disparate educational structures.

2.8.4 Integration of educational and skills taxonomies

The integration of educational taxonomies with skills taxonomies is a critical task that aims to align educational outcomes with the competencies required in the labor market. This integration facilitates better curriculum development, career planning, and workforce development by ensuring that educational programs are directly linked to the skills needed by employers.

Kuodytė, Petkevičius et al. [28] propose a methodology for mapping educational content to job market skills using hierarchical classification and transformer neural networks, specifically tailored for handling complex educational data. The study employs hierarchical classification to manage the complex structure of educational data, where programs and courses are organized into multi-level taxonomies. This allows for the mapping of education data at different granularity levels, ensuring a detailed alignment with job skills. By utilizing transformer-based models, the research effectively transforms textual descriptions of educational programs into embeddings, capturing nuanced relationships between education and skills. This approach leverages the power of modern NLP techniques to handle the intricate connections in educational data. The paper addresses common challenges in education-to-skill mapping, such as imbalanced datasets, complex labeling, and the hierarchical nature of educational structures. These are tackled through advanced neural network algorithms, which enhance the accuracy of the mapping process. The proposed model is demonstrated on national-level data from Lithuania, showcasing its potential for large-scale, practical applications in policy-making, scenario forecasting, and human resource management.

Rentzsch,Staneva et al. [44] review the application of classification systems for job-relevant skills and competencies, comparing traditional, expert-curated taxonomies with modern, data-driven ontologies. The authors examine the role of these classification systems in skills matching (aligning job seekers with suitable jobs) and skills intelligence (labor market analysis and forecasting). The authors discuss curated ontologies and data-driven ontologies, as

a dynamic systems that update in real-time using data analytics and NLP, making them more adaptable to the evolving labor market. The paper emphasizes the future importance of annotating educational programs with skills taxonomies, closely aligning with the thesis focus on integrating education and skills taxonomies through NLP.

In this thesis, we explore the integration of educational taxonomies with dynamic skills taxonomies. Incorporating advanced NLP models like BERT can enhance the precision and scalability of this mapping, ensuring that educational programs are continuously aligned with evolving labor market needs.

3 Data

To successfully generate an education taxonomy, it is crucial to gather comprehensive and authentic data from various educational systems, particularly within Europe. These systems vary not only in structure and qualification types but also operate in different native languages, necessitating the creation of multilingual datasets for accurate representation and classification.

English serves as a vital intermediary language, supporting the standardization of concepts and allowing cross-border comparisons, particularly important for global applications such as Randstad's job matching platforms designed to support job seekers.

As detailed in Section 2, the structural diversity of national education systems—including general, vocational, and specialized pathways—further necessitates a flexible and inclusive approach to data collection to ensure that all qualification types are adequately captured.

From Randstad's perspective, the collection of educational data is driven by the need to populate their applications [41] with accurate and detailed information about education diplomas.

In the applicant tracking system used by Randstad, consultants need to indicate the level, education, and specialization on job orders. To facilitate this, the taxonomy must encompass a wide range of educational diplomas and qualifications, allowing consultants to select from an exhaustive list that matches the real-world educational landscape. The more diplomas we collect, the more accurate and tailored these job requirements can be.

On the Randstad website and job boards, job seekers need the option to add their education from pre-defined lists, which necessitates the inclusion of a broad array of educational programs and diplomas to cover all potential user entries. Talents can indicate whether they have achieved a diploma for their selected education, making it essential to include detailed records of different diploma types and their educational pathways. Additionally, talents must be able to select the institution where they obtained their education, requiring the collection of comprehensive data on educational institutions to ensure accurate representation and selection.

So, the widest possible range of educational diplomas allows for precise matching of job requirements to candidates' qualifications and supports a seamless user experience in documenting and verifying educational backgrounds.

3.1 Data sources

The data was collected from various European countries in line with Randstad's requirements. This data includes educational qualifications, diplomas, and institution details, sourced from national education databases, government publications, and institutional websites. The multilingual nature of the data ensures that the taxonomy accurately reflects the diverse educational systems across Europe, providing the necessary breadth to support Randstad's platforms.

3.1.1 Diplomas from The Netherlands

The Netherlands boasts a well-structured and highly regarded education system, characterized by its emphasis on both academic and vocational pathways. To accurately categorize and map Dutch higher education and vocational qualifications within a comprehensive education taxonomy, it is essential to rely on authoritative sources that provide detailed information about accredited programs, learning outcomes, and competency frameworks. Two key resources for this purpose are the Register Instellingen en Opleidingen (RIO) and the Collaboration Organization for Vocational Education and Training (S-bb.nl).

1. **RIO**

   RIO [18], formerly known as CROHO (Centraal Register Opleidingen Hoger Onderwijs), is the central registry for higher education programs in the Netherlands, managed by DUO (Dienst Uitvoering Onderwijs). This resource provides detailed information about accredited higher education institutions and their programs, including program names (diplomas), levels, and associated learning outcomes. Extracting this

data is crucial for accurately categorizing Dutch higher education diplomas and mapping their qualifications within the education taxonomy.

2. **S-bb.nl**

   S-bb.nl [45] (Samenwerkingsorganisatie Beroepsonderwijs Bedrijfsleven) serves as a hub for information on vocational education and training (VET) in the Netherlands, connecting educational institutions, businesses, and students.

   S-bb.nl provides data on:

   - Vocational education programs and qualifications.
   - Competency frameworks and skill requirements for various professions.
   - Partnerships between educational institutions and industry stakeholders.

   The data from S-bb.nl is particularly valuable for understanding the landscape of vocational education and the skills required in the labor market. This information can be used to enrich the taxonomy by incorporating vocational training pathways and competency-based classifications.

3.1.2 Diplomas from Belgium

Belgium's education system is characterized by its regional diversity, with distinct frameworks governing the Dutch-speaking Flanders region, the French-speaking Wallonia region, and the German-speaking community. This section focuses on the Flanders region, where two key resources—Onderwijskiezer.be and AHOVOKS—provide comprehensive and structured data on educational pathways, qualifications, and competencies. These resources are invaluable for accurately categorizing and mapping Belgian diplomas within the proposed education taxonomy.

1. **Onderwijskiezer.be**

   Onderwijskiezer.be [37] is a comprehensive online platform that provides information and guidance on educational pathways in Flanders (Dutch

part), Belgium. It is designed to help students, parents, and educators make informed decisions about education and career choices. The platform covers all levels of education, from primary to higher education, as well as vocational training and adult education.

Onderwijskiezer.be offers structured data on:

- Educational programs and study options at different levels (primary, secondary, higher education, and vocational training).
- Detailed descriptions of study fields, including learning outcomes, career prospects, and required competencies.
- Tools for comparing educational programs and identifying suitable pathways based on individual interests and abilities.

The platform's structured and detailed data makes it a valuable resource for building an education taxonomy. Its focus on educational pathways, learning outcomes, and career alignment provides a rich foundation for AI models to classify and map educational programs.

2. **AHOVOKS**

AHOVOKS [2] is the Flemish government agency responsible for higher education, adult education, qualifications, and study grants. It plays a crucial role in shaping and regulating the Flemish education system by ensuring quality, promoting accessibility, and supporting students through financial aid programs.

AHOVOKS provides structured data on:

- Accredited vocational education institutions and programs in Flanders.
- Frameworks for qualifications and competencies, including the Flemish Qualifications Framework (VLQ).

AHOVOKS publishes with us a dataset on vocational education in the Flemish region, classified under ISCED [49]. These datasets offer valuable information on vocational programs, including levels and learning outcomes, making them an excellent resource for experimenting with our models and refining the education taxonomy.

3.1.3 Diplomas from Italy

For Italy, the data used to build the education taxonomy comes from Randstad's internal resources, specifically the data collected through their R-One and W-One applications. The Randstad expert team has curated a comprehensive dataset by gathering educational qualifications and programs entered by consultants and talents within these platforms. This dataset includes crucial information such as program names (diplomas), levels, and learning outcomes, which are essential for creating a detailed and accurate taxonomy for the Italian education system.

In R-One, consultants manually enter educational qualifications when creating job orders. This input includes details about the level and specialization of education required for various roles. Over time, this input has provided a rich dataset of educational qualifications relevant to the Italian labor market.

In W-One, talents add their educational background, specifying the diplomas they have achieved, along with the institutions they attended. This information adds another layer of depth to the dataset, offering insights into the educational landscape from the perspective of job seekers.

The dataset curated from R-One and W-One is highly valuable as it reflects real-world applications and expectations in the Italian job market.

3.1.4 Diplomas from Sweden

For Sweden, the educational data is sourced from the Swedish National Agency for Education ('Skolverket') through their publicly available API [22]. This API provides comprehensive datasets on Swedish education programs, including details about program names (diplomas), levels, and learning outcomes.

The API from the Swedish National Agency for Education allows for automated and structured access to up-to-date educational information. This includes data on various types of educational programs, covering everything from primary education to vocational training and higher education.

The API provides a wide range of data across different educational levels

and types, ensuring that the taxonomy includes all relevant Swedish qualifications. Using the API allows for regular updates to the dataset, ensuring that the education taxonomy remains current with any changes in the Swedish education system. The structured format of the API data aids in the seamless integration into the education taxonomy, supporting the classification and mapping processes necessary for Randstad's applications.

3.1.5 Lightcast skills

To integrate skills taxonomy with education taxonomy, we collected structured skills data from the Lightcast API [31], a widely used source for labor market intelligence. The Lightcast Skills Taxonomy (v9.20) provides a hierarchical classification of skills based on real-time labor market analysis, helping to standardize the connection between education and workforce requirements.

While the Lightcast API contains various metadata fields, we focused on the most relevant structured fields necessary for taxonomy alignment:

- **Skill Name:** The official name of the skill as classified by Lightcast.
- **Category:** The broad classification under which the skill falls (e.g., "Technology", "Healthcare").
- **Sub-Category:** A more specific classification within the category (e.g., "Data Science" under "Technology").
- **Description:** A textual explanation of what the skill entails, providing context for classification.

From the Lightcast API, we extracted approximately **33,620** skills, covering a diverse range of industries and domains. This dataset forms the basis for aligning educational qualifications with in-demand skills, enabling more accurate skill-to-education mapping in our research. Samples from the Lightcast Skills Taxonomy (v9.20) are provided in the Appendix A.1.

3.1.6 ESCO

The European Skills, Competences, Qualifications, and Occupations (ESCO v1.2.0) [20] dataset is a comprehensive multilingual classification system developed by the European Commission. It provides a structured framework for identifying and describing occupations, skills, competences, and qualifications across the European Union. The dataset is designed to facilitate labor market transparency, improve job matching, and support workforce development by establishing a common language for skills and occupations.

ESCO organizes its data into many datasets, in this study we focus on these following three:

1. **Occupation Dataset:** The Occupation Dataset provides a structured and standardized representation of occupations, including their definitions, classifications, and associated metadata. This dataset is essential for understanding the roles and responsibilities of various jobs within the labor market. We focus on these key fields:
    - preferredLabel: The primary or most commonly used name for the occupation.
    - description: A detailed description of the occupation, including its tasks and responsibilities.
2. **Skills dataset:** The skills dataset provides a comprehensive taxonomy of skills, competences, and knowledge areas relevant to various occupations. It serves as a foundation for understanding the skill requirements of jobs and supports skill-based matching in the labor market. We focus on these key fields:
    - skillType: Specifies the type of skill (e.g., "transversal", "occupation-specific", "knowledge").
    - preferredLabel: The primary or most commonly used name for the skill.
    - description: A detailed description of the skill, including its relevance and application.

3. **Occupation-Skill relation dataset:** The Occupation-Skill Relation Dataset establishes connections between occupations and the skills required to perform them. This dataset is critical for understanding the skill demands of specific jobs and for aligning education and training with labor market needs. We focus on these key fields:

    - occupationURL: The URI of the occupation, linking it to the corresponding record in the Occupation Dataset.
    - relationType: Specifies the type of relationship between the occupation and the skill (e.g., "essential", "optional").
    - skillType: Indicates the type of skill (e.g., "transversal", "occupation-specific", "knowledge").
    - skillURL: The URI of the skill, linking it to the corresponding record in the Skills Dataset.

This dataset enables the mapping of skill requirements to occupations, supporting applications such as job matching, workforce planning, and curriculum development. Samples from the ESCO Skills Taxonomy (v1.2.0) are provided in the Appendix A.2.

3.2 Pre-processing

3.2.1 ISCED Data

As we discussed before, the structure of ISCED data 2.3, UNESCO designed three hierarchical levels to classify educational programs: Broad Field, Narrow Field, and Detailed Field. In the detailed fields, as Table 3.1 shows, subjects are classified based on UNESCO guidelines, providing a global standard for educational programs.

Table 3.1: Classification of Arts and Humanities Fields According to ISCED.

Level	**Code**	**Category**
Broad Field	02	Arts and humanities
Narrow Field	021	Arts
Detailed Field	0213	Fine arts
Subjects		Art theory Calligraphy Etching Fine art printmaking History of art Painting Philosophy of art Sculpture

We construct a dataset by extracting subjects from the detailed fields of the ISCED taxonomy and labeling them with their corresponding detailed field categories. This process resulted in a structured dataset comprising approximately 1,075 records, with each record containing a subject and its classification. The dataset spans a total of 80 unique labels, representing the detailed field categories used in the ISCED hierarchy.

Table 3.2: Statistical Overview of the ISCED Dataset

Dataset	**Count**
Subjects	1,075
labels	80

As Figure 3.1 shows, the dataset has a large number of categories with an uneven distribution. For fine-tuning a BERT model for a classification task, a hundred examples per category can already be enough. Additionally, the uneven distribution of records across labels further complicates training, as it can lead to bias in model predictions for underrepresented categories.

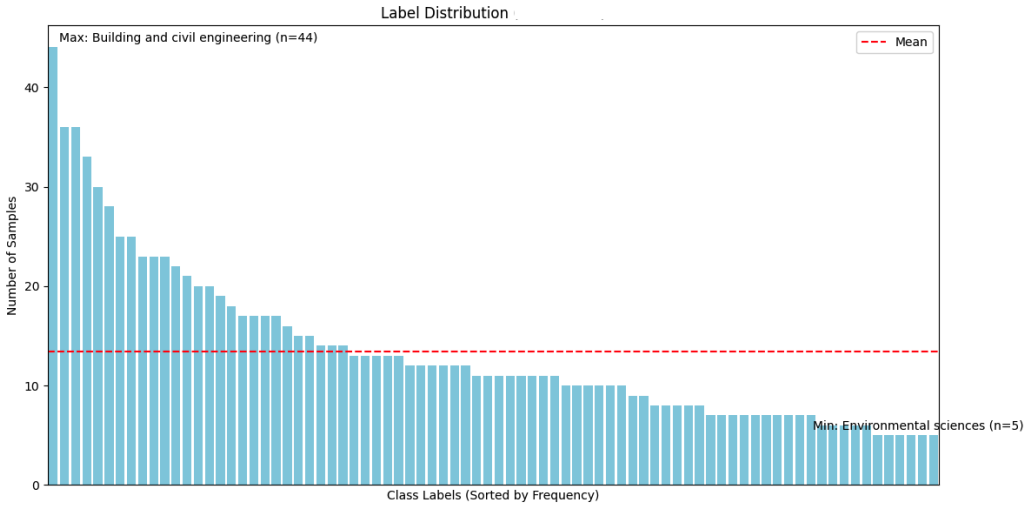

Figure 3.1: Labels distribution in the ISCED-F2013 dataset prior to augmentation.

3.2.2 Educational programs data

To create a unified education taxonomy that covers multiple countries, it was essential to standardize the data collected from various national sources, including Dutch, Belgian, Italian, and Swedish educational programs. Each country's educational data was structured differently, requiring us to format all the collected data into a consistent structure for effective integration and analysis.

To achieve this, we utilized the Pandas framework in Python to clean, normalize and reformat datasets originating from various sources such as CSV files, Excel sheets, and scraped web downloads.

The educational programs from each country were reformatted into a uniform dataset with the following fields shown in Table 3.3.

Table 3.3: Overview of the Key Fields in the Educational Data Collected from Various Countries.

Field	**Description**
Diploma Name (Original Language)	The name of the diploma in its original language, as provided by the source country's education system.
Diploma Name (English)	The translated name of the diploma in English
Level	The educational level of the diploma (e.g., PhD, Master, Bachelor, Vocational).
EQF	The corresponding level on the European Qualifications Framework (EQF)
Short Description (Optional)	A brief description of the diploma, where available, to provide additional context about the educational program.

This standardized dataset ensures consistency and compatibility across different national educational systems, allowing us to integrate these programs into a comprehensive taxonomy that can be used for further processing and modeling.

After formatting the data into this standardized structure, we aggregated the data for each country to assess the scope and coverage of educational programs included in our taxonomy. Below is a summary table 3.4 of the number of educational programs collected from each country.

Table 3.4: Summary of Educational Programs Collected from National Sources in Selected European Countries

Country	**National Source**	**Number of Educational Programs**
Netherlands	RIO	3,775
	S-bb	1,075
Belgium	Onderwijskiezer.be	1,135
	AHOVOKS	559
Italy	Randstad's Database	435
Sweden	Skolverket	65

4 Methods

In this section, we present the methodologies employed for education taxonomy generation and its integration with the skills taxonomy. We begin by exploring multi-class text classification approaches used for structuring the education taxonomy. Next, we outline the data augmentation techniques applied to expand the training dataset. We then describe the methods used to match educational programs with relevant skills from the skills taxonomy. Finally, we introduce the evaluation metrics used to assess the effectiveness and accuracy of our approach.

4.1 Multi-class text classification for education taxonomy generation

To generate the education taxonomy, we formulate the problem as a multi-class text classification task, where each educational program (diploma) is assigned to one of the detailed fields in the ISCED taxonomy (see Figure 2.2). Given that ISCED provides a hierarchical classification of education programs, our goal was to train a model that classifies new educational programs based on their textual descriptions. Educational programs are primarily described through textual titles (e.g., "Bachelor in Mechanical Engineering"). Since ISCED fields also include textual definitions of disciplines, NLP-based classification is a natural choice for mapping titles to ISCED fields.

4.1.1 Dataset preparation

We first prepare the ISCED-labeled dataset (see Section 3.2), which originally contained 1,075 records (i.e. educational subject) classified into 80 categories (i.e. ISCED detailed field).

Encoding the labels Before training our multi-class text classification model, we transformed the ISCED detailed fields into numerical labels using label encoding. This step is essential for ensuring compatibility with BERT-based models, which require numerical target values. Encoding standardizes

label representation, optimizes computational efficiency, and allows the softmax output layer to classify input text into one of 80 ISCED categories. It also enables faster lookups and seamless integration with deep learning frameworks. During inference, predicted numerical labels are mapped back to their corresponding ISCED categories for interpretability.

Train and validation split To ensure a balanced representation of each class in both the training and validation sets, we applied a stratified split based on class labels. Given that the dataset exhibits class imbalance, this approach preserves the original label distribution, preventing the model from becoming biased toward more frequent categories while ensuring sufficient representation of minority classes. We allocate **85%** of the data for training and **15%** for validation, maintaining consistency by setting a fixed random state. This stratified sampling strategy helps the model generalize more effectively, ensuring that both the training and validation sets accurately reflect the real-world distribution of educational categories in the dataset.

Data augmentation As shown in Figure 3.1, the label distribution exhibits significant class imbalance, with several low-frequency labels, we address both class imbalance and data scarcity through a two-stage augmentation strategy. First, we reduce the imbalance by oversampling low-frequency labels using synonym replacement and WordNet-based substitutions [33]. Subsequently, we apply NLPAug [32] to the balanced dataset obtained from the first step to expand the dataset significantly while preserving the overall label distribution and semantic integrity of the input. This library offers a wide range of augmentation techniques, and in our case, we leverage:

- **Back translation:** We translated each subject in the ISCED list 3.2.1 into another language (such as: French, German, and Spanish) and then back to the original language (English), creating paraphrased versions and enhancing dataset diversity while retaining the original meaning.
- **Synonym replacement:** By substituting words in the subject content with their synonyms, we generated multiple variations of the same record, further enriching the dataset.

These augmentation methods (as Figure 4.1 shows) increased our dataset size

by approximately **five** times, enhancing its diversity and improving imbalance labels.

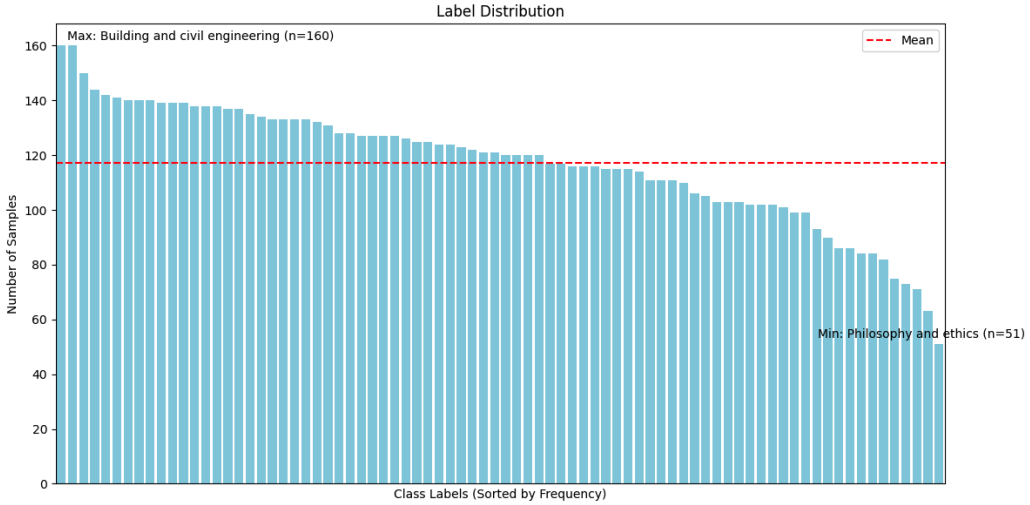

Figure 4.1: Labels distribution in the ISCED-F2013 dataset after augmentation.

4.1.2 BertTokenizer and encoding the data

Tokenization is the process of transforming raw text into smaller linguistic units (tokens), which are then converted into numerical representations suitable for machine learning models. Since BERT operates on tokenized text, we employ BERT's WordPiece tokenizer, which efficiently handles out-of-vocabulary words by breaking them into subword units.

To ensure compatibility with BERT's input format, we utilize a pre-trained BERT tokenizer, aligning the tokenization process with the model's architecture. Tokenization and encoding are applied separately to both the training and validation datasets using the `batch_encode_plus` function. The key parameters used in this process include:

- `add_special_tokens=True`: Ensures that BERT-specific special tokens (`[CLS]` and `[SEP]`) are included, marking the beginning and end of sequences.

`[CLS]` Educational Subject `[SEP]`

- `return_attention_mask=True`:Generates an attention mask, distinguishing actual tokens from padding tokens, allowing the model to focus only on meaningful inputs.
- Padding to a fixed length: All sequences are padded to a specified maximum length (256 tokens in this case) to maintain consistency across batches.
- `return_tensors='pt'`: Converts the encoded data into PyTorch tensors, making them suitable for training in a PyTorch-based deep learning framework.

Following tokenization and encoding, the dataset is further structured into three key components:

- Input IDs – *Numerical representations of tokenized* text.
- Attention Masks – *Binary masks indicating which tokens should be attended to.*
- Labels – *Encoded target categories (ISCED detailed fields).*

These processed inputs form the final training and validation datasets, which are used for fine-tuning the BERT model in the classification task.

4.1.3 BERT Pre-trained Model

In this study, each educational program title is treated as a distinct sequence, where each sequence is classified into one of 80 predefined ISCED detailed fields. We compare multiple pre-trained BERT models to evaluate their effectiveness in this classification task:

1. **bert-base-uncased:** A widely used English BERT model, fine-tuned on large English corpora.

2. **mBERT (Multilingual BERT):** A variant trained on 104 languages, enabling classification of non-English diploma titles directly.
3. **modernBERT:** A more recent transformer-based model optimized for multi-lingual and domain-specific applications.

For each model, we define the `num_labels` parameter to specify the number of classification categories (80 ISCED labels). Since attention outputs and hidden states are not required for this task, we disable `output_attentions` and `output_hidden_states` to optimize memory usage and computational efficiency.

To efficiently handle the dataset during training, we utilize the DataLoader class, which enables batched processing and integrates a sampling strategy to optimize data selection.

- Training Set: We employ a RandomSampler, which ensures diverse data selection by randomly shuffling samples at each epoch, helping the model generalize better.
- Validation Set: A SequentialSampler is used to maintain the original order of validation samples, ensuring consistency in performance evaluation.

Given memory constraints in our computing environment, we set the batch size to 3, striking a balance between model performance and hardware limitations.

For model optimization, we employ an optimizer that iterates over the trainable model parameters, updating their weights based on computed gradients. Key hyperparameters, such as learning rate (`lr`) and epsilon (`eps`), are carefully configured to ensure numerical stability.

Through empirical experimentation, we determined that training for five epochs (`epochs=5`) provides an optimal trade-off between convergence and model performance. To enhance training stability and prevent overfitting, we implement a learning rate scheduler with a linear decay strategy:

- During an initial warmup period, the learning rate increases linearly from zero to a predefined value, ensuring a smooth transition into optimization.
- After reaching the peak learning rate, it gradually decays to zero over the remaining epochs, preventing sudden parameter updates that could destabilize training.

4.2 Matching education taxonomy with relevant skills

The effectiveness of matching educational taxonomies with relevant skills largely depends on the availability, structure, and granularity of the data used in the process. The diversity in qualification frameworks, skill taxonomies, and occupational classifications presents a significant challenge in establishing a coherent and scalable mapping between educational programs and labor market competencies.

As outlined in Section 3, our datasets consist of structured qualifications frameworks NLQF 3.1.1 and skill taxonomies from sources such as ESCO 3.1.6 and Lightcast 3.1.5. The goal is to create an initial dataset that accurately links qualifications to relevant skills, which serves as training material for a transformer-based model capable of predicting the skills associated with any given educational program.

To construct this qualification-to-skill mapping, we propose and implement two distinct approaches, named entity recognition (NER) for skill extraction and embedding-based similarity matching. By implementing these approaches, we aim to get initial dataset enables us to fine-tune BERT in a multi-label classification setting, allowing for automated skill prediction for any educational program.

Figure 4.2 presents a high-level comparison of the two proposed approaches, in the following sections, we will detail these approaches and their application in solving this problem effectively.

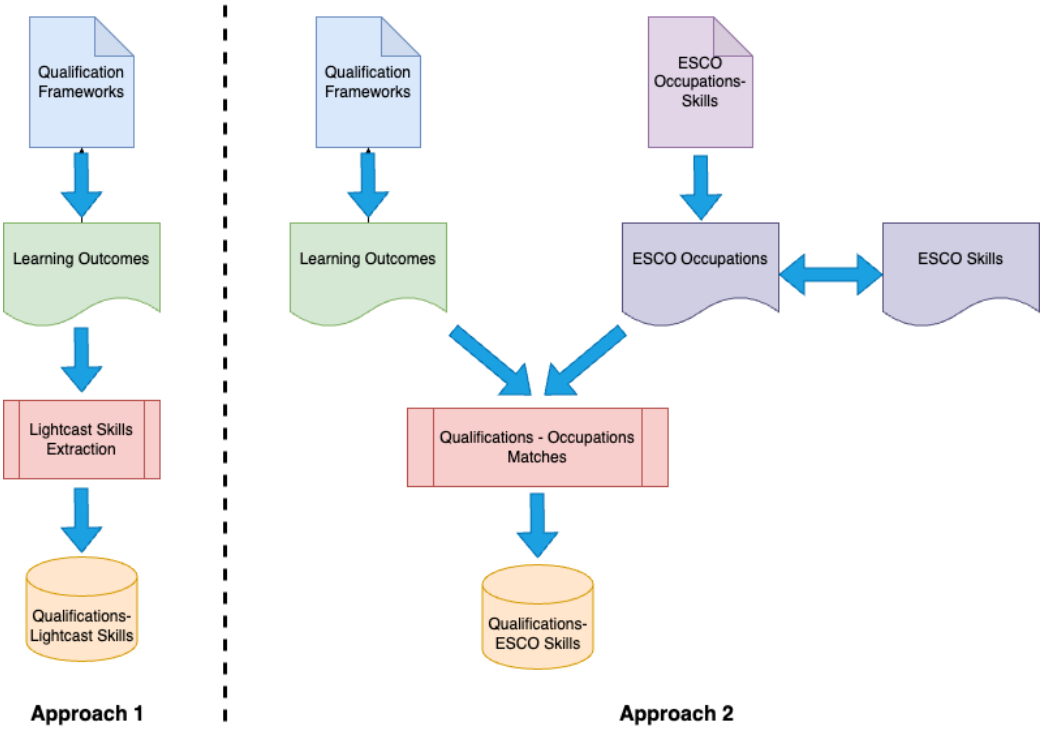

Figure 4.2: High-level comparison of the two approaches used for matching Education taxonomy with skills.

4.2.1 Approach 1: Named Entity Recognition (NER) for skill extraction

In this approach, we leverage NER to extract relevant skills directly from the learning outcomes of Dutch qualifications frameworks. Learning outcomes provide a detailed description of the competencies, knowledge, and abilities acquired through a qualification. Since these outcomes inherently contain skill-related information, an NER-based approach enables automated identification of skills within the text. Figure 4.3 illustrates the pipeline of this approach, from inputting the learning outcomes to extracting structured skill entities using the NER model.

Figure 4.3: Approach 1: NER-Based Education-to-Skills Matching Pipeline.

4.2.1.1 Dataset

To implement NER-Based extraction, we rely on two primary datasets:

- The Dutch Qualifications Framework (NLQF) dataset (3.1.1) provides a structured list of qualification titles along with their corresponding learning outcomes. The dataset is available in both English and Dutch, in this approach, we use the English version.
- The Lightcast skills dataset 3.1.5, which contains a comprehensive collection of skills, each associated with: *Skill Name* (the standard name of the skill), *Skill description* (a textual explanation of the skill's scope and relevance) and *Skill categories* (the broader classification of the skill, grouping related competencies)

4.2.1.2 SkillNER module

SkillNER [3] is a skill extraction library built on top of the spaCy [26] NLP framework. It utilizes a pre-defined skill ontology and a phrase-matching mechanism to identify skill-related terms in textual data. SkillNER relies on a lightcast skill database that contains skill identifiers $skill\_id$, skill names, and associated metadata. This ontology serves as the foundation for identifying relevant skills in text.

It employs a PhraseMatcher from spaCy to detect skill mentions in text. The matcher compares n-grams (sequences of words) in the input text against the skill ontology, ensuring accurate and efficient skill detection.

SkillNER assigns a confidence score to each extracted skill, indicating the likelihood that the detected phrase corresponds to a valid skill. This score is derived from the similarity between the detected phrase and the skill ontology.

SkillNER provides structured output in the form of annotations, which include:

- $skill\_id$: A unique identifier for the extracted skill.
- doc_node_value: The skill name or phrase detected in the text.
- $score$: A confidence score ranging from 0 to 1, indicating the reliability of the extraction.
- $match\_type$: The type of match (e.g., $full\_matches$ for exact matches or $ngram\_scored$ for partial matches).

In this study, SkillNER was applied to extract skills from learning outcomes associated with qualifications. The process involved the following steps:

- Data Preparation: A dataset containing qualification titles and their corresponding learning outcomes was preprocessed to ensure consistency and remove noise (e.g., special characters, stopwords).
- Skill Extraction: SkillNER was applied to each learning outcome to identify skill-related phrases. The module processed the text and returned a list of extracted skills, along with their confidence scores and match types.
- Post-Processing: The extracted skills were filtered based on their confidence scores (e.g., retaining only skills with a score above a predefined threshold). Additionally, duplicate skills were removed to ensure a clean and concise output.
- Structured Output: The final output was organized into a structured format, including the qualification title, extracted skills, and their associated metadata ($skill\_id, score$). This structured data will be then used in our next step, multi-label classification.

Next, SkillNER output will be integrated with transformers to enhance its contextual understanding and improve performance on complex texts.

4.2.2 Approach 2: Education to ESCO Occupation-Skills Alignment

In this approach, we match educational programs with relevant skills using a different strategy. Since ESCO occupations are already mapped to specific skills, we can establish an indirect link between education and skills by first matching qualification frameworks NLQF to occupations and then assigning skills based on these occupations. Linking qualifications to occupations provides a structured and scalable approach to assigning skills to educational programs, effectively bridging the gap between education taxonomies and relevant skills.

Once the qualification-to-occupation mapping is established, we fine-tune BERT to predict skills for new educational programs based on learned patterns. This approach ensures that even educations without predefined mappings can be dynamically associated with relevant occupations and skills. Figure 4.4 illustrates the pipeline of this approach, from inputting the learning outcomes and ESCO dataset to matching skills.

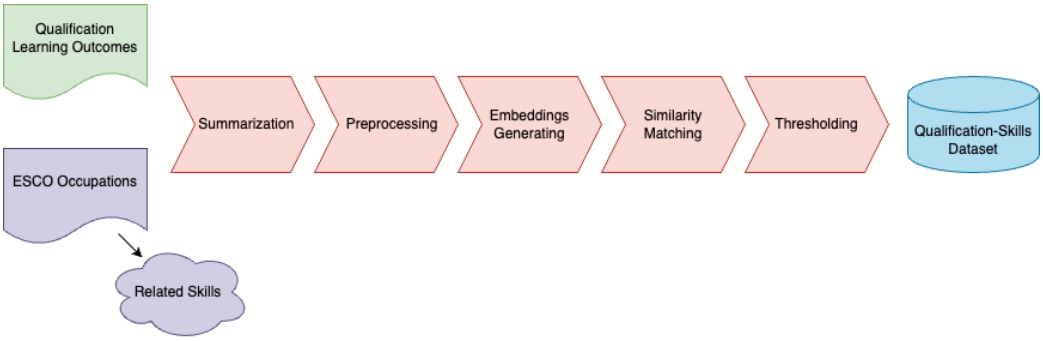

Figure 4.4: Approach 2: Embedding-based pipeline for matching educational programs to skills via ESCO occupations.

4.2.2.1 Datasets

In this approach, we rely on two key datasets to establish the link between educational programs, occupations, and skills: the ESCO occupation-skills dataset and the NLQF qualifications dataset. Each dataset provides structured

information that enables us to build an automated education-to-skills matching model.

The **ESCO** (3.1.6) dataset serves as a comprehensive labor market classification system. While ESCO contains various types of structured data, we specifically focus on its occupation-to-skills relationships.

- Occupations: ESCO provides a structured list of occupational titles each mapped to industry-specific skills.
- Occupation description, a structured description of the occupations.
- Skills Taxonomy: Each occupation is linked to a set of relevant skills.

By using these relationships, we can infer which skills are required for specific occupations and, by extension, which skills are relevant to educational programs when mapped to occupations.

The Netherlands Qualifications Framework (**NLQF**) (3.1.1), as described in previous section, provides a structured representation of educational programs, including their level, classification, and competencies. NLQF provides detailed learning outcomes, making it a rich source for qualification-to-occupation alignment. We primarily focus on:

- Program Tilte (The official title of the educational qualification).
- Learning Outcomes (A structured description of the competencies, skills, and knowledge gained from the program).

4.2.2.2 Summarization

Learning outcomes and occupation descriptions in both the NLQF and ESCO datasets are often verbose, containing detailed information that may introduce redundancy and unnecessary complexity when performing similarity computations. Given that embedding-based models perform optimally on concise and semantically rich text, we apply text summarization techniques to generate more compact representations of both learning outcomes and occupation

descriptions while preserving their key semantic information. To achieve this, we employ two state-of-the-art transformer-based models: T5 (Text-To-Text Transfer Transformer) [38] for learning outcomes and BART (Bidirectional and Auto-Regressive Transformers) [29] for occupation descriptions.

T5 is a versatile transformer-based model that frames all natural language processing (NLP) tasks as a text-to-text problem. It uses a unified architecture where both the input and output are represented as text strings. For summarization, T5 is fine-tuned on datasets like CNN/DailyMail, enabling it to generate high-quality summaries. Learning outcomes are typically long and structured, making them well-suited for T5's text-to-text framework. The T5-base variant strikes a balance between model size and performance, making it computationally efficient for large datasets.

Learning outcomes were preprocessed to remove unnecessary formatting and standardized into a consistent structure. The T5-base model was used, with a maximum input length of 512 tokens and a maximum output length of 150 tokens. Parameters such as $num\_beams = 4$, $length\_penalty = 2.0$, and $early\_stopping = True$ were used to optimize summary quality. Learning outcomes were processed in batches to improve efficiency, with a batch size of 32 to balance speed and GPU memory usage.

BART is a sequence-to-sequence model that combines a bidirectional encoder (like BERT) and an auto-regressive decoder (like GPT). It is pre-trained on a denoising objective, where it learns to reconstruct corrupted text, making it highly effective for text generation tasks like summarization. Occupation descriptions are often more complex than learning outcomes, requiring a model with a larger context window. BART's 1024-token input limit accommodates longer texts, ensuring that key information is not truncated. Its abstractive capabilities produce fluent and coherent summaries, which are essential for matching occupation descriptions to learning outcomes.

Occupation descriptions were also cleaned and standardized to ensure consistency. The text was tokenized using BART's tokenizer, which supports a maximum input length of 1024 tokens. The $BART - large - cnn$ model was used, fine-tuned on the CNN/DailyMail dataset for summarization. The maximum output length was set to 150 tokens, with parameters like $num\_beams = 4$ and $length\_penalty = 2.0$ to enhance summary quality.

Occupation descriptions were processed in batches with a batch size of 16 to accommodate the larger model size and GPU memory constraints.

Summarizing these descriptions offers several advantages, it improves text representation by generating more precise embeddings and reducing noise in similarity matching, increases matching efficiency by accelerating computational performance without losing essential details, and enhances generalization by allowing models to focus on key attributes, ultimately improving the accuracy of alignment between qualification frameworks and relevant occupations.

4.2.2.3 SBERT as embedding-based model

Since both learning outcomes and occupation descriptions describe what a learner is expected to achieve and what an employee is expected to perform, they provide a rich semantic basis for embedding-based alignment.

To establish meaningful connections between qualification frameworks and occupations, we leverage Sentence-BERT (SBERT) [43] to generate dense, context-aware vector representations of both summarized learning outcomes and occupation descriptions. Specifically, we use the $all\_mpnet\_base\_v2$ model, a state-of-the-art SBERT variant built on the MPNet architecture, which integrates masked and permuted language modeling to produce embeddings with rich semantic depth. A key improvement of SBERT over traditional BERT that SBERT employs a bi-encoder architecture that allows independent encoding of input texts and enables efficient similarity computation through approximate nearest neighbor search. This architectural design makes it significantly faster and more scalable for large-scale matching tasks. The $all\_mpnet\_base\_v2$ model is particularly well-suited for our task, as it supports multilingual input and eliminates the need for manual pooling strategies, generating sentence-level embeddings in a single step. This makes it ideal for aligning educational qualifications with occupations and their associated skill profiles in multilingual, high-dimensional data environments.

Using the $all\_mpnet\_base\_v2$ model, we generated embeddings for both texts. The model captured the semantic similarity between the learning outcome and the occupation description, despite differences in phrasing. For instance, the terms "problem-solving skills in software development" and

"design, code, and test software applications" were mapped closely in the embedding space, reflecting their shared focus on software development and system design. Similarly, "modern programming languages" and "programming languages and frameworks" were recognized as semantically equivalent.

The final sentence embeddings are stored as 768-dimensional vectors in a shared semantic space, where similar concepts are mapped closer together. These embeddings serve as the foundation for the next stage of processing, where we compute semantic similarity to establish the best matches between qualifications and occupations.

4.2.2.4 Similarity matching

Once dense vector representations of summarized learning outcomes and the summarized descriptions of ESCO occupations are generated using BERT, the next step is to measure their semantic similarity to determine the most relevant occupation for each educational program. Since learning outcomes describe the competencies gained through education, and occupation descriptions outline required job competencies, semantic similarity enables an evidence-based mapping between qualifications and labor market needs.

To achieve this, we apply cosine similarity, a widely used metric for measuring the angular similarity between high-dimensional embeddings. Unlike direct keyword-based matching, cosine similarity allows us to compare contextual meaning, making it particularly effective for competency-based alignment. Given two normalized embeddings, A (learning outcome embedding) and B (occupation description embedding), cosine similarity is computed as equation:

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

where:

- $A \cdot B$ represents the dot product of the two vectors.
- $\|A\|$ and $\|B\|$ are the L2 norms (magnitudes) of the vectors.

- The resulting similarity score ranges from -1 (completely opposite) to 1 (identical), with 0 indicating no similarity.

Similarity threshold To ensure the accuracy and relevance of the matched skills, we apply a thresholding to filter out weak matches and prioritize the most relevant occupation for each qualification framework. Since cosine similarity assigns a score between -1 and 1, where higher values indicate greater similarity, it is crucial to define an optimal threshold to distinguish meaningful matches from irrelevant ones.

A common similarity threshold for text-matching tasks using cosine similarity typically ranges between 0.6 and 0.8 (Achananuparp et al., 2008; Zhu et al., 2010) [1]. For domain-specific tasks like taxonomy alignment, a stricter threshold of 0.7 – 0.85 ensures high-quality matches while minimizing false positives (Rajpal & Rathore, 2014) [40].

Since learning outcomes and occupation descriptions may not always share exact wording, a threshold between 0.65 and 0.8 is recommended.

Mathematically, a match is considered valid if:

$$\text{cosine\_similarity}(A, B) \geq 0.7$$

where A represents the educational program embedding and B represents the skill embedding.

Once the education-to-occupation relationships are established, the next step involves inferring relevant skills by leveraging ESCO's occupation-to-skills mappings.

4.2.3 Fine-tuning BERT for automatic skill prediction

Given a dataset of qualifications and their associated skills, the goal is to develop a machine learning model that can predict relevant skills for new qualification titles, handle the multi-label nature of the problem, where each

qualification title can be associated with multiple skills, and generalize well to unseen data, even when some skills are underrepresented in the dataset.

We formulate the problem as a multi-label text classification task, where the input is a qualification title (text) and the output is a set of binary labels indicating the presence or absence of each skill. The methodology consists of the following steps:

1. **Model initialization:** The BERT-based model (BertForSequenceClassification) was initialized with pre-trained weights from the bert-base-uncased variant. The output layer was configured to predict binary labels for each skill, with the number of output units equal to the total number of unique skills in the dataset.

2. **Data preprocessing:** The dataset was preprocessed to group skills by qualification titles and encode skills as binary labels using multi-label binarization. The dataset was split into training and test sets using an $80 - 20$ split ratio. Qualification titles were tokenized using the BERT tokenizer, with a maximum sequence length of 128 tokens. Padding and truncation were applied to ensure uniform input sizes. Skills were encoded as binary vectors using multi-label binarization, where each skill was represented as a binary label (1 if present, 0 otherwise).

3. **Loss Function and Optimizer:** The Binary Cross-Entropy Loss with Logits $BCEWithLogitsLoss$ was used as the loss function, suitable for multi-label classification tasks. The $AdamW$ optimizer was employed with a learning rate of 2×10^{-5} and weight decay of 0.01 to prevent overfitting.

4. **Data augmentation:** To address the limited size of the dataset, we applied text augmentation techniques on the training set, including: **Synonym replacement:** Replacing words with their synonyms to introduce variability. **Back-translation:** Translating text to another language and back to the original language to generate paraphrased versions.

5. **Training loop:** The model was trained for 10 epochs with a batch size of 16. During each epoch, the following steps were performed, **Forward Pass:** Input tokens and attention masks were passed through

the model to obtain logits. **Loss Calculation:** The loss was computed by comparing the predicted logits with the ground truth binary labels. **Backward Pass:** Gradients were computed using backpropagation. **Optimization:** Model parameters were updated using the AdamW optimizer. Training loss was monitored after each epoch to ensure the model was learning effectively.

6. **Evaluation:** After each epoch, the model was evaluated on the test set using precision, recall, and F1 score as metrics. A prediction threshold of 0.5 was used to convert logits into binary predictions, balancing the trade-off between precision and recall.

4.3 Evaluation metrics

To assess the effectiveness of our education taxonomy classification model, we employ a range of evaluation metrics that provide both overall performance insights and class-level analysis. Given the multi-class nature of the task and the presence of imbalanced class distributions, we utilize metrics that evaluate general accuracy, class-specific performance, and misclassification patterns.

The Macro F1-score calculates the F1-score for each class individually and then averages the scores across all classes. This ensures that each class contributes equally to the final metric, regardless of its frequency in the dataset.

$$\text{Macro F1} = \frac{1}{N} \sum_{i=1}^{N} \text{F1-score}_i \tag{1}$$

where N is the number of classes.

The Micro F1-score aggregates true positives (TP), false positives (FP), and false negatives (FN) across all classes before computing a single F1-score. Unlike Macro F1, which treats all classes equally, Micro F1 is weighted by class frequency.

$$\text{Micro F1} = \frac{2 \times \sum TP}{2 \times \sum TP + \sum FP + \sum FN} \tag{2}$$

The Weighted F1-score computes the F1-score for each class and then weights it by the number of instances in that class. This metric helps balance the impact of large and small classes.

$$\text{Weighted F1} = \sum_{i=1}^{N} w_i \times \text{F1-score}_i \tag{3}$$

where w_i is the proportion of class i in the dataset.

Precision and Recall are fundamental metrics in classification tasks:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

Precision measures how many of the predicted positive instances are actually correct, while Recall measures how many actual positive instances were correctly identified.

The classification report provides a detailed breakdown of Precision, Recall, and F1-score for each taxonomy category. It is particularly useful in identifying well-performing classes versus those needing improvement.

A confusion matrix visually represents the misclassification patterns by showing the number of times each class was correctly or incorrectly predicted. Given the large number of labels (80 classes), we focus on the top misclassified labels.

$$\mathbf{CM} = \begin{bmatrix} TP_{1,1} & FP_{1,2} & FP_{1,3} & \dots & FP_{1,N} \\ FN_{2,1} & TP_{2,2} & FP_{2,3} & \dots & FP_{2,N} \\ FN_{3,1} & FN_{3,2} & TP_{3,3} & \dots & FP_{3,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ FN_{N,1} & FN_{N,2} & FN_{N,3} & \dots & TP_{N,N} \end{bmatrix} \tag{6}$$

Each entry $TP_{i,i}$ represents correctly classified instances for class i, while FP and FN represent misclassified instances.

5 Experiments and results

In this chapter, we present the experiments conducted to evaluate the effectiveness of our proposed methods. A total of three experiments were performed, each corresponding to one of the three core research questions outlined in this study.

5.1 Experiment 1: Education taxonomy generation

In this experiment, we evaluate and compare the performance of four transformer-based models: BERT (bert-base-uncased), ModernBERT (ModernBERT-base), mBERT (bert-base-multilingual-uncased), and XLM-R (xlm-roberta-large) in the task of education taxonomy classification. The models were fine-tuned on the ISCED dataset consisting of 1,075 subject titles mapped to 80 ISCED detailed fields (see Section 3.2.1). These models were selected based on their capabilities in handling monolingual and multilingual text data, which is critical for accurately categorizing educational programs across different languages and regions.

The objective of this experiment is to assess the efficiency, accuracy, and multilingual adaptability of these transformer models in automating the generation of an education taxonomy. Through this evaluation, we address our **(RQ1)**: "*How can transformers improve the efficiency, accuracy, and multilingual adaptability of education taxonomy generation compared to traditional methods?*"

To answer this, we analyze model performance across multiple evaluation metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis. By comparing the classification effectiveness of monolingual (BERT, ModernBERT) and multilingual (mBERT, XLM-R) models, we investigate the extent to which transformer models can enhance taxonomy generation compared to rule-based and manual classification approaches.

Table 5.1 shows the performance comparison of four transformer models -BERT, ModernBERT, mBERT, and XLM-R - on the education taxonomy classification task. The results indicate that mBERT achieves the highest

overall performance, with the best scores in accuracy (0.864), macro F1-score (0.871), and micro F1-score (0.874). These results highlight the model's strong multilingual capabilities, which are particularly beneficial for cross-lingual education data. ModernBERT follows closely, outperforming the standard BERT model across all metrics, including macro F1-score (0.835) and overall F1-score (0.830), making it a competitive alternative with improved classification capability. While BERT still demonstrates solid performance for monolingual tasks, achieving a macro F1-score of 0.820, it lags slightly behind ModernBERT and mBERT. XLM-R underperforms relative to the other models, particularly in accuracy (0.751) and F1-score (0.744), indicating limitations in its ability to handle fine-grained classification in this domain.

Table 5.1: Performance comparison of transformer models in education taxonomy classification

Model	Accuracy	Precision	Recall	F1-score	Macro F1-score	Micro F1-score
BERT	0.816	0.829	0.816	0.813	0.820	0.816
ModernBERT	0.831	0.845	0.831	0.830	0.835	0.831
mBERT	**0.864**	**0.875**	**0.863**	**0.861**	**0.871**	**0.874**
XLM-R	0.751	0.757	0.753	0.744	0.779	0.753

Table 5.2 provides a detailed classification report, including macro and weighted averages for precision, recall, and F1-score. The report confirms that mBERT outperforms the other models, particularly in handling class imbalances. BERT also maintains high precision and recall, making it a reliable option for English-based taxonomies. However, XLM-R struggles to correctly classify certain categories, especially in low-resource taxonomy labels, as indicated by its lower macro F1-score.

Table 5.2: Detailed classification report for mBERT, including macro and weighted Precision, Recall, and F1-Score in education taxonomy classification.

Class	Precision	Recall	F1-Score	Support
Building and civil engineering	0.67	0.52	0.59	23
Crop and livestock production	0.61	0.83	0.70	23
Teacher training with subject specialisation	0.95	0.95	0.95	22
Computer use	0.95	0.95	0.95	22
Electricity and energy	0.62	0.62	0.62	21
Audio-visual techniques and media production	0.65	0.62	0.63	21
Database and network design and administration	0.70	0.76	0.73	21
Teacher training without subject specialisation	0.82	0.86	0.84	21
Motor vehicles, ships and aircraft	0.72	0.62	0.67	21
Electronics and automation	0.61	0.52	0.56	21
Work skills	0.57	0.81	0.67	21
Secretarial and office work	0.43	0.62	0.51	21
Medical diagnostic and treatment technology	0.74	0.81	0.77	21
Nursing and midwifery	0.70	0.67	0.68	21
Materials (glass, paper, plastic and wood)	0.62	0.62	0.62	21
Basic programmes and qualifications	0.95	0.95	0.95	21
Medicine	0.76	0.62	0.68	21
Language acquisition	0.89	0.85	0.87	20
Chemical engineering and processes	0.70	0.70	0.70	20
Mining and extraction	0.84	0.80	0.82	20
Occupational health and safety	0.55	0.60	0.57	20
Environmental protection technology	0.75	0.75	0.75	20
Transport services	0.65	0.55	0.59	20
Textiles (clothes, footwear and leather)	0.86	0.60	0.71	20
Child care and youth services	0.86	0.90	0.88	20
Personal skills and development	0.65	0.68	0.67	19
Food processing	0.68	0.89	0.77	19
Software and applications development and analysis	0.65	0.68	0.67	19
Fashion, interior and industrial design	0.67	0.74	0.70	19
Management and administration	0.68	0.68	0.68	19
Mechanics and metal trades	0.54	0.74	0.62	19
Veterinary	1.00	1.00	1.00	19
Care of the elderly and of disabled adults	0.83	0.79	0.81	19
Social work and counselling	0.60	0.79	0.68	19
Military and defence	0.89	0.84	0.86	19
History and archaeology	0.71	0.79	0.75	19
Handicrafts	0.83	0.28	0.42	18
Statistics	0.82	0.78	0.80	18
Travel, tourism and leisure	0.94	0.83	0.88	18
Training for pre-school teachers	1.00	1.00	1.00	18

Class	Precision	Recall	F1-Score	Support
Hotel, restaurants and catering	0.73	0.89	0.80	18
Forestry	0.94	0.89	0.91	18
Natural environments and wildlife	1.00	0.94	0.97	18
Architecture and town planning	0.80	0.67	0.73	18
Fisheries	0.93	0.78	0.85	18
Law	0.76	0.72	0.74	18
Literature and linguistics	0.77	0.59	0.67	17
Library, information and archival studies	0.88	0.88	0.88	17
Earth sciences	0.78	0.82	0.80	17
Finance, banking and insurance	0.73	0.65	0.69	17
Sociology and cultural studies	0.82	0.82	0.82	17
Music and performing arts	0.71	0.88	0.79	17
Protection of persons and property	0.77	0.59	0.67	17
Domestic services	0.70	0.82	0.76	17
Sports	0.87	0.76	0.81	17
Dental studies	0.84	0.94	0.89	17
Education science	0.82	0.88	0.85	16
Fine arts	0.92	0.75	0.83	16
Hair and beauty services	0.71	0.62	0.67	16
Traditional and complementary medicine and therapy	0.81	0.87	0.84	15
Wholesale and retail sales	0.65	0.87	0.74	15
Horticulture	1.00	0.73	0.85	15
Community sanitation	0.75	0.80	0.77	15
Biochemistry	0.67	0.67	0.67	15
Political sciences and civics	0.85	0.73	0.79	15
Therapy and rehabilitation	0.86	0.80	0.83	15
Marketing and advertising	0.88	0.47	0.61	15
Physics	0.53	0.60	0.56	15
Religion and theology	0.93	0.93	0.93	14
Journalism and reporting	0.71	0.92	0.80	13
Mathematics	1.00	0.85	0.92	13
Literacy and numeracy	0.92	0.92	0.92	13
Environmental sciences	0.85	0.85	0.85	13
Economics	0.85	0.85	0.85	13
Biology	0.80	0.67	0.73	12
Chemistry	0.83	0.91	0.87	11
Psychology	0.85	1.00	0.92	11
Pharmacy	1.00	1.00	1.00	11
Accounting and taxation	0.80	0.44	0.57	9
Philosophy and ethics	0.73	1.00	0.84	8
Accuracy			0.86	1406
Macro avg	0.88	0.87	0.87	1406
Weighted avg	0.87	0.86	0.86	1406

Figure 5.1 presents the confusion matrix for the top misclassified categories using the mBERT model, highlighting common misclassification patterns among education taxonomy labels. The diagonal values, representing correct predictions indicating variability in class-wise accuracy. This range suggests the presence of class imbalance, where certain classes are underrepresented in the dataset, leading to fewer correct predictions for those classes. Off-diagonal cells, which represent misclassifications, are predominantly 0, with only a few instances of 1. This indicates that the model performs well overall, with minimal confusion between classes. However, the occasional misclassifications (values of 1) highlight specific pairs of classes that may share semantic or contextual similarities, posing a challenge for the model.

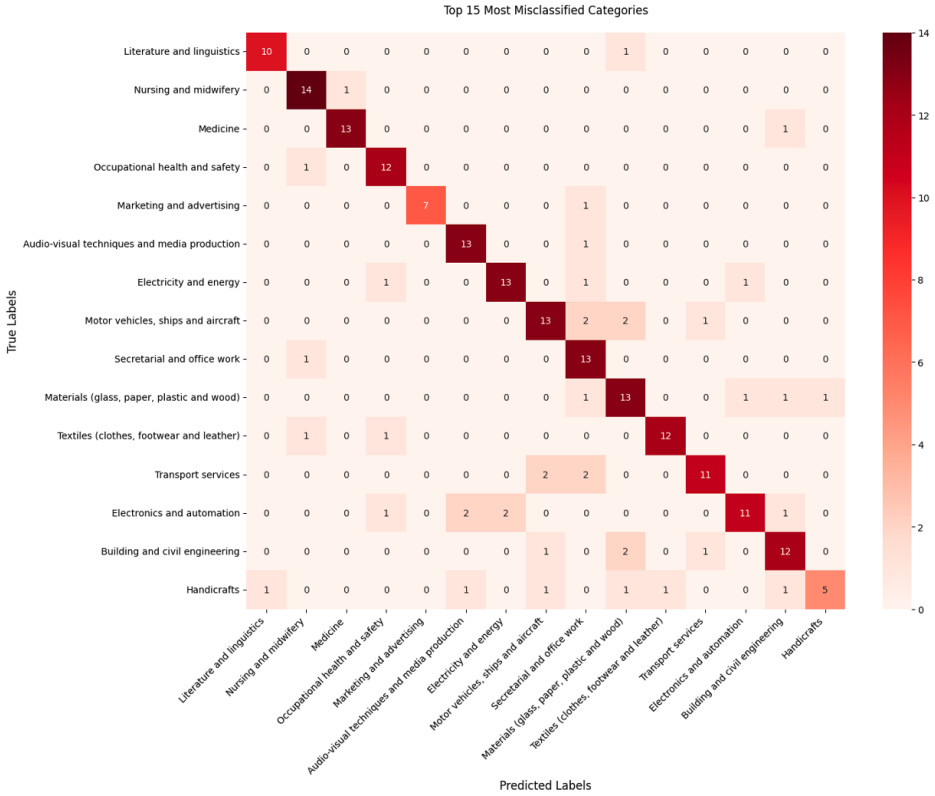

Figure 5.1: Confusion Matrix of the Top Misclassified Categories Using the mBERT Model in Education Taxonomy Classification.

The results indicate that mBERT outperforms the other models in terms of macro-averaged performance metrics, making it the best-suited model for multilingual education taxonomy classification. However, BERT and modernBERT also show competitive performance, especially when working with English-based datasets. While XLM-R was expected to perform well on multilingual data, its lower accuracy and higher misclassification rates suggest that it may require further fine-tuning or domain-specific adaptation for education taxonomy tasks.

5.2 Experiment 2: Generalization

This experiment aims to assess the performance of our fine tuned BERT model in classifying the education programs from different countries (Netherlands, Belgium and Italy), using taxonomy labels assigned by human experts as the ground truth. The results of this evaluation provide insights into whether machine generated taxonomies align with human classifications, or if it would need further refinement. The findings from this experiment contribute to answering our **(RQ2)**: "How can an education taxonomy be standardized and generalized across different countries and institutions to ensure cross-border comparability?"

Table 5.3 demonstrates the varying levels of generalization performance across the Netherlands, Belgium and Italy. In the Netherlands, the BERT model achieves good alignment with human-assigned taxonomy labels, indicating high accuracy and validating the model's ability to effectively learn and classify within its training context. In Belgium, the BERT model shows moderate alignment with human classifications, indicating a reasonable ability to generalize across different education systems. In contrast, the results for Italy appear lower in accuracy and F1-score. However, these misclassifications may not necessarily indicate poor model generalization.

Instead, experts feedback from Randstad suggest that the predicted taxonomy labels often outperform the human-assigned classifications. This indicates that discrepancies between the two may arise due to inconsistencies or limitations in human labeling rather than actual model errors. To illustrate this, Figure 5.2 highlights patterns where the model systematically assigns

taxonomies that are semantically more appropriate than those provided by human experts.

Table 5.3: Performance Comparison of BERT-Based Taxonomy Classification on Netherlands, Belgium and Italy Educational Data

Metric (Weighted)	Netherlands Data	Belgium Data	Italy Data
Accuracy	0.864	0.617	0.406
Precision	0.875	0.674	0.467
Recall	0.863	0.617	0.406
F1-Score	0.861	0.619	0.409

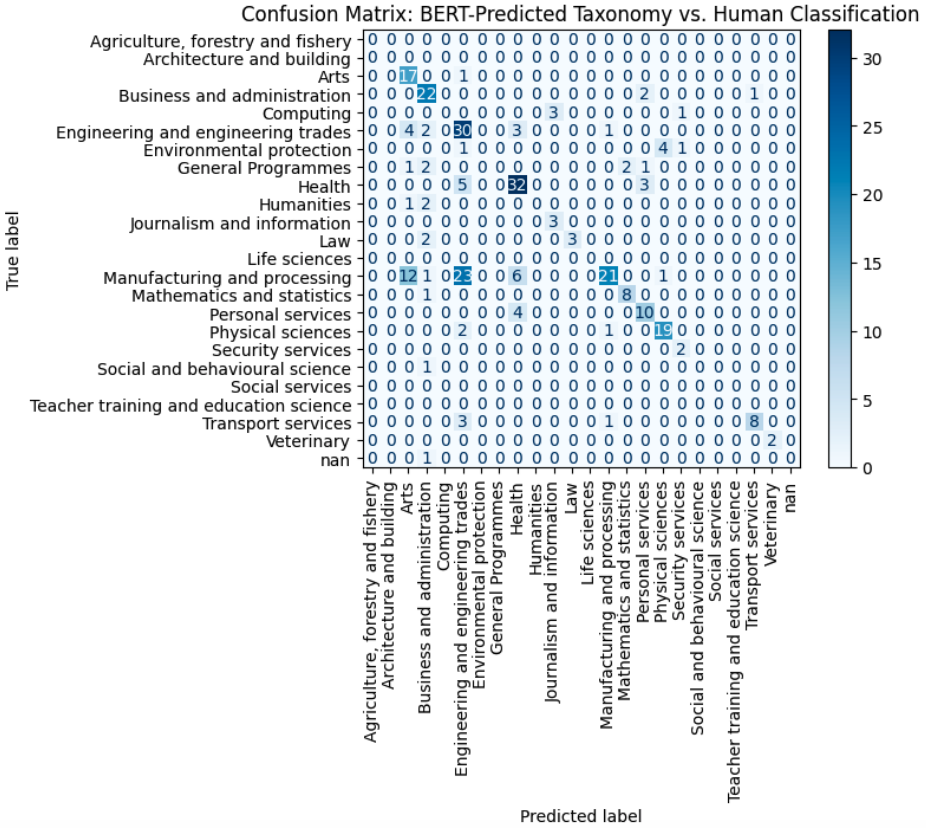

Figure 5.2: Confusion Matrix of BERT-Predicted Taxonomy vs. Human Classification on the Italy data.

Table 5.4 presents concrete examples where BERT correctly places a qualification into a more relevant taxonomy than the expert label, further supporting the hypothesis that the model offers a more structured and accurate classification approach. For instance, the qualification "Fashion Designer" is classified by human experts under "Manufacturing and Processing", whereas BERT assigns it to "Arts," which is a more logical categorization given the nature of the field. Similarly, "Computer Sciences and Technologies" is classified by human experts under "Computing," while BERT assigns it to "Information and Communication Technologies (ICTs)," which provides a broader and more precise classification.

Table 5.4: Misclassified examples in the Italian dataset – comparison between BERT-predicted and Human-assigned taxonomies.

Diploma Title	**Human Label**	**BERT Label**
Dental technician operator	Manufacturing and processing	Health
Biological chemical operator	Manufacturing and processing	Biological and related sciences
Fashion designer	Manufacturing and processing	Arts
Dietology	Personal services	Health
Linguistic mediator	Humanities	Languages
Computer sciences and technologies	Computing	Information and Communication Technologies (ICTs)
Aviation construction expert	Manufacturing and processing	Engineering and engineering trades

5.3 Experiment 3: Matching education taxonomy with relative skills

In the following sections, we present the experiments conducted for key steps and techniques within both approaches we introduced before. We explore how these techniques can be leveraged to establish meaningful connections between educational programs and relevant skills, addressing a critical gap in aligning education systems with labor market demands. By systematically evaluating the effectiveness of these methods, we aim to answering our **(RQ3)**: "*How closely are the education and skills taxonomies related and how can we establish a link between them using NLP techniques?*"

5.3.1 NER

In this experiment, we evaluate the effectiveness of Named Entity Recognition (NER) in extracting relevant skills from the learning outcomes of educational qualifications. The goal is to determine how well the extracted skills align with a standardized skills taxonomy and whether the assigned confidence scores from the SkillNER tool can be used as a reliable measure of relevance.

To extract skills from the learning outcomes of educational qualifications, we employed SkillNER (see section 4.2.1.2), a domain-specific NER tool optimized for skill recognition. The dataset used in this experiment consists of structured qualifications from the NLQF framework, where each qualification includes its title and a set of learning outcomes that describe the expected competencies. The SkillNER pipeline was applied to these learning outcomes, extracting a list of potential skills along with their assigned confidence scores.

The extracted skills were categorized into two groups based on their match type:

1. Full Matches (score = 1) → Direct matches to skills in the lightcast taxonomy, considered high confidence.
2. N-gram Scored Matches ($0 <$ score < 1) → Partial matches, requiring further evaluation to determine relevance.

To assess the effectiveness of this approach, we conducted two key evaluations, confidence Score Analysis, examining the distribution of confidence scores between relevant and irrelevant skills. And ranking-based, retrieval-oriented evaluation metrics including Precision@5, Recall@5, and NDCG@5, measuring how well SkillNER ranks relevant skills higher than irrelevant ones.

A key aspect of the evaluation involved analyzing the distribution of confidence scores assigned by SkillNER. The confidence score histogram (Figure 5.3) shows that most of the extracted skills are relevant (in green), receiving scores within the defined threshold range of 0.7 to 1, while a smaller subset (in red) was identified as irrelevant. This indicates that SkillNER assigns higher confidence scores to genuinely relevant skills, reinforcing the effectiveness of 0.7 as an optimal threshold for filtering out irrelevant extractions while preserving meaningful skill mappings.

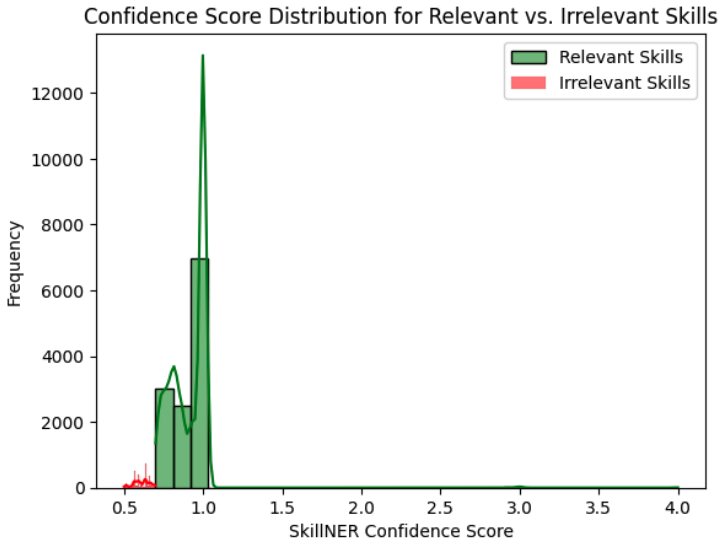

Figure 5.3: Distribution of Confidence Scores for Relevant and Irrelevant Skills Extracted by SkillNER.

As Table 5.5 shows, the results demonstrate strong performance: **NDCG@5** reached **0.991**, indicating that SkillNER ranks relevant skills almost perfectly at the top of the list. In addition, we obtained a **Precision@5** of **0.929** and a **Recall@5** of **0.714**, suggesting that the majority of top-ranked skills are relevant and that a substantial portion of all relevant skills are successfully

captured within the top five positions.

Table 5.5: Performance Metrics for SkillNER Skill Extraction (Top-5 Skills)

Metric	**Value**
Precision@5	0.929
Recall@5	0.714
NDCG@5	0.991

The results demonstrate that Named Entity Recognition (NER) using SkillNER is an effective approach for extracting relevant skills from educational qualification learning outcomes. By leveraging confidence scores and ranking metrics, we were able to filter out irrelevant extractions and prioritize highly relevant skills.

5.3.2 Summarization

Although we initially explored the application of abstractive summarization models such as BART and T5 to compress lengthy descriptions of skills, occupations, and qualification learning outcomes, we ultimately decided not to include a full evaluation of this step. The primary reason was the absence of reliable human-written reference summaries, which are necessary for accurate and meaningful evaluation using standard metrics such as ROUGE. Without ground-truth references, ROUGE scores – which rely on n-gram overlap with human summaries – cannot provide valid indicators of quality. As a result, while summarization was used as a preprocessing step to reduce verbosity and noise in the text data for downstream tasks (e.g., embedding and similarity matching), we did not perform a quantitative evaluation of its performance. Future work could address this gap by incorporating human-annotated summaries or expert-curated references for more robust evaluation.

5.3.3 Multi-label Classification

This section presents a comparative evaluation of our approaches for the multi-label skill classification task. The skills datasets contain around (Lightcast: 33,620 and ESCO: 13,492) unique skills, with each qualification linked to an average of 25 skills. This high-dimensional label space adds substantial complexity to the classification problem, highlighting the need for robust models capable of handling sparse and imbalanced output distributions. We assess the performance of both approaches using identical qualification title datasets to enable direct benchmarking. Three state-of-the-art transformer encoders – BERT, DistilBERT, and DeBERTa – are fine-tuned for each approach to examine model-agnostic effectiveness. The evaluation employs precision, recall and F1-score metrics, with particular attention to handling class imbalance inherent in skill prediction tasks. This systematic comparison aims to identify optimal combinations of methodology and architecture for educational skill mapping applications.

Table 5.6 presents the comparative performance of three transformer architectures based on the approach 1. The results reveal significant variations in model effectiveness, with DeBERTa achieving the highest F1-score followed closely by DistilBERT, while BERT substantially underperforms.

Table 5.6: Multi-label classification results for Approach 1 (Lightcast skills).

Model	Precision	Recall	F1 Score
BERT	0.255	0.382	0.306
DistilBERT	0.428	0.659	0.519
DeBERTa	0.433	0.669	**0.521**

To gain further insight into model performance at the skill level, Tables 5.7 and 5.8 present the top five best and worst performing skills predicted by the DeBERTa model in Approach 1, ranked by their individual Precision, Recall, and F1-scores. These results highlight both the strengths and limitations of the model in classifying specific skill categories.

The superior performance of DeBERTa can be attributed to its disentangled attention mechanism, which appears particularly adept at capturing

Table 5.7: Top-5 best predicted skills by DeBERTa in approach 1 based on Precision, Recall, and F1-Score.

Skill label	Precision	Recall	F1 Score
TL 9000 Standard	1.000	1.000	1.000
Arithmetic	1.000	1.000	1.000
Point Of Sale	1.000	0.900	0.947
Management Training And Development	1.000	0.833	0.909
Corporate Governance Of ICT	1.000	0.812	0.896

Table 5.8: Top-5 worst predicted skills by DeBERTa in approach 1 based on Precision, Recall, and F1-Score.

Skill label	Precision	Recall	F1 Score
Prototyping	0.20	0.10	0.14
Protein Secondary Structure	0.25	0.11	0.15
Public Announcement	0.33	0.20	0.25
Prototype (Manufacturing)	0.40	0.35	0.37
Prototype (Computer Science)	0.45	0.39	0.42

the nuanced relationships between qualification titles and their associated skills. Second, DistilBERT's competitive results (within 0.5% of DeBERTa's F1-score) despite its reduced size demonstrate that model distillation effectively preserves predictive capability while improving computational efficiency. Third, BERT's comparatively weak performance (41% lower F1-score than DeBERTa) highlights the limitations of baseline transformer architectures for this specific task.

The results for Approach 2, which predicts ESCO skills through embedding-based matching with occupational profiles, reveal distinct performance patterns (Table 5.9). DeBERTa again emerges as the top performer (F1: 0.586), demonstrating improvement over BERT's baseline (F1: 0.175), with DistilBERT maintaining competitive results (F1: 0.531). This substantial performance gap underscores the critical role of model architecture in handling skill-occupation relationships.

Also for approach 2, to go further insight into model performance at the skill

Table 5.9: Multi-label classification results for Approach 2 (ESCO skills)

Model	Precision	Recall	F1 Score
BERT	0.212	0.121	0.175
DistilBERT	0.448	0.648	0.531
DeBERTa	0.475	0.692	**0.586**

level, Tables 5.10 and 5.11 present the top five best and worst performing skills predicted by the DeBERTa model, ranked by their individual Precision, Recall, and F1-scores. These results highlight both the strengths and limitations of the model in classifying specific skill categories.

Table 5.10: Top-5 best predicted skills by DeBERTa in approach 2 based on Precision, Recall, and F1-Score.

Skill label	Precision	Recall	F1 Score
deal with challenging people	1.000	1.000	1.000
apply person-centred care	0.771	0.870	0.818
assess students	0.717	0.933	0.811
assist students in their learning	0.717	0.933	0.811
legal requirements in the social sector	0.724	0.913	0.807

Table 5.11: Top-5 worst predicted skills by DeBERTa in approach 2 based on Precision, Recall, and F1-Score.

Skill label	Precision	Recall	F1 Score
organise rehearsals	0.22	0.20	0.21
organise resources for the vehicle showroom	0.23	0.21	0.22
organise relapse prevention	0.25	0.23	0.24
organise resources for artistic production	0.29	0.27	0.28
organise technical operating information for vehicles	0.32	0.28	0.30

The relative performance ranking (DeBERTa > DistilBERT > BERT) mirrors Approach 1, suggesting model capabilities generalize across different skill ontologies. However, Approach 2 shows wider performance dispersion, with DeBERTa achieving 12.3% higher F1 than in Approach 1, while BERT drops

by 43%. This indicates that ESCO skill prediction benefits disproportionately from advanced architectures.

Second, the precision-recall results show consistent patterns, DeBERTa achieves the best balance (precision: 0.475, recall: 0.692). DistilBERT shows marginally lower precision (0.448) with comparable recall (0.648). BERT fails catastrophically on recall (0.121), suggesting inadequate representation learning.

Third, the absolute performance differences between approaches highlight methodological impacts:

- Embedding-based ESCO matching (Approach 2) yields higher peak F1 (0.586 vs 0.5216)
- NER-extracted Lightcast skills (Approach 1) produce more stable baseline performance

These results carry important implications for educational knowledge graph construction. The superior performance of embedding-based approaches with advanced models suggests that occupation-skill relationships in ESCO provide richer signal for qualification alignment than direct skill extraction from learning outcomes. However, DistilBERT's robust performance (90% of DeBERTa's F1) confirms that efficient models can effectively leverage this signal, enabling scalable deployments.

6 Discussion

In this section, we provide a deeper analysis of the experimental results. We further interpret key findings, explore patterns observed in the model's performance, and assess the effectiveness of the proposed approaches. Additionally, we discuss the limitations of our work.

6.1 Interpretations

The results from the **first** experiment provide valuable insights into the effectiveness of transformer-based models for education taxonomy classification. Among the three evaluated models—BERT, mBERT, and XLM-R—mBERT demonstrated the highest overall performance, highlighting the advantages of multilingual contextual learning in accurately classifying educational programs. This suggests that handling linguistic variations and cross-lingual semantics is crucial for improving taxonomy assignments, particularly when working with diverse educational datasets.

The monolingual BERT model (bert-base-uncased) also performed well, achieving competitive classification metrics comparable to mBERT. This indicates that for taxonomies operating within a single language, monolingual models remain an effective choice. However, the lower performance of XLM-R (xlm-roberta-large) suggests that while it excels in general multilingual tasks, it may face challenges in fine-grained educational classification, possibly due to differences in pretraining objectives and tokenization strategies.

These findings emphasize the importance of selecting a transformer model suited to the linguistic diversity and structural complexity of the taxonomy being developed. They also suggest that future improvements could focus on fine-tuning multilingual models with domain-specific educational data to enhance classification precision.

The **second** experiment provides critical insights into the generalization ability of our fine-tuned BERT model when applied to education programs from different countries (Belgium and Italy). The results indicate that while the model aligns well with human-assigned taxonomies in Belgium, its performance

in Italy appears lower in terms of accuracy and F1-score. However, upon closer examination, these discrepancies do not necessarily reflect poor model performance but rather highlight inconsistencies in human labeling.

A detailed error analysis reveals that BERT often assigns taxonomies that are semantically more appropriate than those provided by human experts. In several cases, the model produces taxonomies that align more logically (based on the feedback by Randstad's experts) with the actual field of study, suggesting that strict evaluation against human classifications may underestimate the model's effectiveness. For example, "Fashion Designer" is labeled under "Manufacturing and Processing" by human experts, whereas BERT assigns it to "Arts," which better represents the nature of the qualification.

These findings suggest that the model is capable of learning meaningful patterns from educational data, even when expert classifications contain inconsistencies. This supports the potential for automated taxonomy classification models to assist or refine human-assigned taxonomies, reducing ambiguity and improving consistency across educational systems. Additionally, this highlights the importance of re-evaluating human-assigned taxonomies and integrating AI-driven insights into the classification process to enhance standardization across different countries.

The **third** experiment evaluates the effectiveness of two distinct approaches, NER-based skill extraction (Approach 1) and embedding-based ESCO matching (Approach 2), for aligning educational programs with relevant skills. Additionally, the experiment benchmarks the performance of three state-of-the-art transformer architectures, BERT, DistilBERT, and DeBERTa, to assess their suitability for this task. The results reveal key insights into the strengths and limitations of each approach.

Regarding model performance across approaches, DeBERTa consistently achieves the highest score, indicating superior contextual representation and feature extraction capabilities. DistilBERT performs competitively, suggesting that lightweight models can still capture meaningful relationships while being more computationally efficient. BERT performs poorly on both approaches, probably due to its older architecture and weaker ability to model complex relationships between educational programs and skills.

Approach 2, highlighting that occupation-skill relationships in ESCO provide a richer and more structured signal for qualification alignment. Approach 1 demonstrates more stable baseline performance, making it a reliable option for cases where structured occupation-skill data is unavailable.

These findings have important implications for educational knowledge graph construction. The superior performance of embedding-based approaches with advanced models suggests that leveraging structured taxonomies like ESCO provides more informative and consistent qualification-skill mappings. However, the strong performance of DistilBERT further confirms that efficient transformer models can be effectively deployed at scale, making them practical for real-world applications.

6.2 Limitations

One of the primary limitations of this study is the availability and access to structured educational data. The lack of centralized and standardized datasets for educational programs across different countries introduced challenges in data collection, preprocessing, and taxonomy alignment. Furthermore, the absence of human-assigned taxonomies for all datasets and countries presents a significant challenge. Additionally, the strategy used for assigning taxonomies varies across education systems, as each country follows a distinct structural framework. These differences further complicate the standardization and generalization of educational taxonomies across multiple regions.

Another limitation in matching education to relevant skills is the dependency on existing skills datasets and the lack of comprehensive learning outcomes or descriptions for some educational programs. In both approaches, our methodology was constrained by the availability and structure of the data. Each approach also presents specific limitations, Approach 1 relies on predefined entity lists (e.g., Lightcast skills), making it less adaptable to new or evolving skill taxonomies. Approach 2 depends on structured occupation-skill relationships from ESCO, which may not always be available for other skill taxonomies such as Lightcast. These data constraints influenced the evaluation and effectiveness of both approaches.

Another limitation in this study is the lack of a comprehensive evaluation

of the summarization models. While we used abstractive summarization as a preprocessing step to improve the quality of input data, the absence of human-annotated summaries for evaluation prevented us from performing a rigorous assessment of the summarization's effectiveness. This limited our ability to fully understand the impact of summarization on downstream tasks, such as embedding and similarity matching. Additionally, the reliance on automatic summarization models, which may struggle with domain-specific nuances, introduces a potential risk of information loss or misrepresentation, further complicating the assessment of their quality.

Lastly, we discuss the limitations of the lack of user-based evaluation for matching education with relevant skills. Assessing the quality of these matches requires extensive validation from domain experts, which is highly resource intensive given the large-scale nature of the data. Future research could incorporate expert review processes or crowd-sourced validation to enhance the reliability of education-to-skill mappings.

7 Conclusion

This research explored the use of transformer-based NLP models for education taxonomy generation and matching it with relevant skills taxonomies. Through a series of experiments, we evaluated the effectiveness of different approaches to answer the following research questions:

RQ1: How can transformers improve the efficiency, accuracy, and multilingual adaptability of education taxonomy generation compared to traditional methods? Our results demonstrate that transformer-based models significantly enhance the automation of education taxonomy classification, offering improved accuracy and efficiency over manual or rule-based approaches. The multilingual capabilities of mBERT proved particularly effective in handling education data from different linguistic backgrounds, outperforming monolingual models in cross-border classification tasks. However, results also indicated that pretrained transformer models alone are not sufficient for full standardization, requiring further domain adaptation and fine-tuning on education-specific datasets.

RQ2: How can an education taxonomy be standardized and generalized across different countries and institutions to ensure cross-border comparability? Our findings indicate that the fine-tuned BERT model successfully generalized across different countries and institutions, demonstrating its ability to classify educational programs consistently despite variations in national education structures. The success of this generalization suggests that transformer-based models can be applied effectively to multiple educational systems, contributing to cross-border comparability and standardization.

RQ3: How closely is the education taxonomy related to the skills taxonomy, and how can we establish a link between them using NLP techniques? This study introduced two distinct approaches for mapping education taxonomy to relevant skills: NER-based skill extraction (Approach 1) and embedding-based matching (Approach 2). Results showed that embedding-based methods leveraging structured education-occupation-skill relationships in ESCO produced stronger alignment, while NER-based extraction offered more stable baseline performance but was dependent on

predefined skill lists. As a result, matching the education taxonomy to relevant skills is effective using both approaches, with performance largely dependent on the availability and structure of the data.

7.1 Future Work

Several opportunities exist for improving the methods and expanding the scope of this research:

- Fine-tuning multilingual models with more countries education data to improve taxonomy classification across diverse education systems.
- Finding a suitable validation on the education-skills matching would provide a real-world assessment of education-to-skill alignment quality.
- Explore ontology-based approaches to enhance education-to-skills matching by leveraging structured knowledge representations. Integrating semantic reasoning and hierarchical relationships from educational and occupational ontologies could improve alignment accuracy and interpretability, enabling more context-aware and scalable mappings.

By addressing these challenges, future research can contribute to the development of more standardized, scalable, and globally applicable education taxonomies, facilitating better integration between education systems and labor market demands.

A Appendices

A.1 Lightcast Skills (v9.20) examples

Table A.1: Samples from the Lightcast Skills Taxonomy (v9.20). Source: [31]

Skill Name	Sub-Category	Category	Description
Build Management	IT Management	Information Technology	Build management is a specialized skill that involves coordinating and overseeing the complete build process for software projects. It includes managing source code, dependencies, and build configurations, as well as ensuring that builds are executed correctly and efficiently. Build managers use tools and technologies such as version control systems, continuous integration servers, and build automation tools to streamline the build process and produce high-quality software releases. This role requires strong technical knowledge, project management skills, and problem-solving abilities.
Essential Tremor	Neurology	Health Care	Essential Tremor is a neurological disorder that causes involuntary shaking or tremors primarily in the hands, but can also affect the head, voice, and other areas. It is considered a specialized skill in the medical field as diagnosis and treatment require specific expertise and knowledge.
Paint Tool SAI	Graphic and Visual Design Software	Design	Paint Tool SAI is a digital painting software that is specifically designed for creating and editing digital art. It has a simple and intuitive interface, and offers a variety of tools and features for drawing, painting, and coloring. SAI is favored by many artists due to its ability to create smooth and clean lines, as well as its flexibility in adjusting brush settings and colors.
Drug Regulatory Affairs	Pharmacology and Drug Discovery	Science and Research	Drug Regulatory Affairs is a specialized skill that involves the compilation, submission and follow-up of drug regulatory documentation required by health authorities for drug approval, marketing authorization, and post-marketing surveillance. It requires expertise in pharmaceuticals, drug development, pharmacology, toxicology, and legal and regulatory requirements.
Community Reinvestment Act (CRA) Lending	General Lending	Finance	Community Reinvestment Act (CRA) lending requires specialized skills as it is a federal law that requires banks to provide financial services to low- and moderate-income communities. Banks must adhere to specific requirements and regulations to demonstrate their commitment to serving these communities.
Glomerular Diseases	Nephrology	Health Care	Glomerular diseases are a group of conditions that affect the glomeruli, which are tiny filters in the kidneys that remove waste and excess fluids from the blood. These diseases can cause inflammation and damage to the glomeruli, leading to symptoms such as proteinuria (protein in the urine), hematuria (blood in the urine), and reduced kidney function.
Embedded Value Accounting	Specialized Accounting	Finance	Embedded Value Accounting (EVA) is a specialized accounting technique used in the insurance industry to estimate the long-term value of insurance policies. EVA takes into account future cash flows from policies and calculates the present value of expected profits.
Contract Auditing	Contract Management	Business	Contract auditing is a specialized skill that involves determining whether contracts are being fulfilled according to their terms and conditions. Contract auditors analyze financial records, internal controls, and other data to identify discrepancies and ensure compliance.
Boat Maintenance	Sea and Waterway Transportation	Transportation, Supply Chain, and Logistics	Boat maintenance is a specialized skill that requires knowledge and experience in various areas such as electrical systems, engine maintenance, plumbing, hull and deck maintenance, and safety procedures. Proper boat maintenance is essential to ensure the boat is safe to use.

A.2 ESCO Skills (v1.2.0) examples

Table A.2: Samples from the ESCO Skills Taxonomy (v1.2.0). Source:

Occupation Name	Skill Name	Relation Type	Skill Type
Technical Director	Theatre techniques	Knowledge	Essential
	Organise rehearsals	Skill/Competence	Essential
	Write risk assessment on performing arts production	Skill/Competence	Essential
	Coordinate with creative departments	Skill/Competence	Essential
	Adapt to artists' creative demands	Skill/Competence	Essential
	Negotiate health and safety issues with third parties	Skill/Competence	Essential
	Adapt designers' work to the performance venue	Skill/Competence	Essential
	Promote health and safety	Skill/Competence	Essential
	Coordinate technical teams in artistic productions	Skill/Competence	Essential
	Write technical riders	Skill/Competence	Optional
Precision Device Inspector	Electrical engineering	Knowledge	Optional
	Micromechatronic engineering	Knowledge	Optional
	Microelectronics	Knowledge	Optional
	MOEM	Knowledge	Optional
	Instrument performance elements	Knowledge	Optional
	Waste removal regulations	Knowledge	Optional
	Microprocessors	Knowledge	Optional
	Electronics	Knowledge	Optional
	Mechanical engineering	Knowledge	Optional
	Microoptics	Knowledge	Optional
	Micromechanics	Knowledge	Optional
	Microelectromechanical systems	Knowledge	Optional
	Interpret circuit diagrams	Skill/Competence	Optional
	Use precision tools	Skill/Competence	Optional
Air Traffic Safety Technician	Air transport law	Knowledge	Essential
	Airport safety regulations	Knowledge	Essential
	Common aviation safety regulations	Knowledge	Essential
	Surveillance radars	Knowledge	Essential
	Implement airside safety procedures	Skill/Competence	Essential
	Install electrical and electronic equipment	Skill/Competence	Essential
	Use technical drawing software	Skill/Competence	Essential
	Follow airport safety procedures	Skill/Competence	Essential
	Carry out preventive airport maintenance	Skill/Competence	Essential
	Use testing equipment	Skill/Competence	Essential
	Assist in the conducting of flight checks	Skill/Competence	Essential
	Comply with air traffic control operations	Skill/Competence	Essential
	Implement safety management systems	Skill/Competence	Essential
	Monitor customer safety on apron	Skill/Competence	Essential
	Maintain electronic equipment	Skill/Competence	Essential
	Follow industry codes of practice for aviation safety	Skill/Competence	Essential
	Ensure aircraft compliance with regulation	Skill/Competence	Essential
	Operate radar equipment	Skill/Competence	Essential
	Air traffic management	Knowledge	Optional
	Electrical engineering	Knowledge	Optional
	Aircraft mechanics	Knowledge	Optional
	Types of aircraft	Knowledge	Optional

References

[1] P. Achananuparp, X. Hu, and X. Shen. The evaluation of sentence similarity measures. In *DaWaK 2008, Springer*, 2008.

[2] AHOVOKS. Agency for higher education, adult education, qualifications, and study grants. `https://www.ahovoks.be`, 2024.

[3] Anas Ait. Skillner: A tool for skill extraction from text. `https://github.com/AnasAito/SkillNER`, 2021. Accessed: 2023-10-15.

[4] World Bank. World development report 2019: The changing nature of work, 2019.

[5] James Bessen, Maarten Goos, Anna Salomons, and Wiljan Van den Berge. The dynamics of skill demand: A perspective on the future of work. *Labour Economics*, 70:101968, 2021.

[6] Benjamin S. Bloom. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Longman, 1956.

[7] Sandra Bohlinger. Promoting lifelong learning through the eqf: The role of non-formal and informal learning. *Journal of Education and Work*, 32(4):388–403, 2019.

[8] Burning Glass Technologies. Skills taxonomy: A framework for understanding the labor market. Technical report, Burning Glass Technologies, 2020.

[9] Cedefop. *Qualifications Frameworks: Implementation and Impact*. Publications Office of the European Union, 2017.

[10] Cedefop. Analysis and overview of national qualifications framework developments in european countries. `https://www.cedefop.europa.eu/en/publications/3074`, 2018.

[11] Cedefop. Using learning outcomes to match job seekers with labour market needs. `https://www.cedefop.europa.eu/en/publications/3081`, 2020.

[12] Daniel Cologna, Michael Hart, and James McBride. Integrating skills and education taxonomies: A case study of european labor markets. *Policy Quarterly*, 14(4):23–30, 2018.

[13] European Commission. The european qualifications framework for lifelong learning (eqf), 2008.

[14] European Commission. European qualifications framework: Supporting learning, work, and cross-border mobility. `https://op.europa.eu/en/publication-detail/-/publication/8c973a2b-75e9-11ea-a07e-01aa75ed71a1`, 2020.

[15] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, 2020.

[16] Peter Dahler-Larsen. National qualifications frameworks: What we know and what we do not know. *European Journal of Education*, 53(3):365–376, 2018.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[18] Dienst Uitvoering Onderwijs (DUO). Rio - register van instellingen en opleidingen. `https://www.duo.nl`, 2024. Accessed: 2024-12-27.

[19] European Commission. Eqf implementation report 2020, 2020.

[20] European Commission. Esco: European skills, competences, qualifications, and occupations. `https://ec.europa.eu/esco/portal`, 2023. Retrieved from https://ec.europa.eu/esco/portal.

[21] Eurydice. The structure of the european education systems 2018/19: Schematic diagrams, 2018.

[22] Swedish National Agency for Education. Swedish national agency for education api. `https://susanavet2.skolverket.se/swagger-ui/index.html`, 2024. Accessed: 2024-12-27.

[23] World Economic Forum. The future of jobs report. https://www.weforum.org/reports/the-future-of-jobs-report-2020/, 2020.

[24] Ofelia Garcia, Tove Skutnabb-Kangas, and María Estela Torres-Guzmán. *Imagining Multilingual Schools: Languages in Education and Glocalization*. Multilingual Matters, 2009.

[25] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*, 2021.

[26] Matthew Honnibal and Ines Montani. spacy: Industrial-strength natural language processing in python. https://spacy.io/, 2020. Accessed: 2023-10-15.

[27] Ernesto Jiménez-Ruiz, Jorge Gracia, Nuno Silva, Alireza Aghaebrahimian, and Anna Tordai. A machine learning approach to multilingual and cross-lingual ontology matching. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference*, pages 602–617. Springer, 2018.

[28] Vilija Kuodytė and Linas Petkevičius. Education-to-skill mapping using hierarchical classification and transformer neural network. *Applied Sciences*, 11(13):5868, 2021.

[29] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[30] Lightcast. The lightcast skills taxonomy: Organizing the world of work, 2022.

[31] Lightcast. Lightcast skills taxonomy api. https://www.lightcast.io, 2024. Accessed: 2024-12-27.

[32] Edward Ma. Nlpaug: A python package for data augmentation in nlp. https://github.com/makcedward/nlpaug, 2019. Accessed: 2024-01-10.

[33] George A Miller. Wordnet: a lexical database. *Communications of the ACM*, 1995.

[34] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[35] NLQF. The dutch national qualifications framework (nlqf). Technical report, National Coordination Point NLQF, 2019.

[36] OECD. Education at a glance 2018: Oecd indicators, 2018.

[37] Onderwijskiezer.be. Educational programs in belgium. `https://www.onderwijskiezer.be`, 2024. Accessed: 2024-12-27.

[38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Google, Mountain View, CA 94043, USA, Journal of Machine Learning Research 21 1-67*, 2020.

[40] R. K. Rajpal and Y. Rathore. A novel technique for ranking of documents using semantic similarity. *International Journal of Computer Science*, 2014.

[41] Randstad. Annual report 2024: Empowering individuals and businesses. `https://www.randstad.com/s3fs-media/rscom/public/2025-02/Randstad_Annual_Report_2024_0.pdf`, 2024.

[42] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL)*, pages 147–155, 2009.

[43] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019.

[44] Mila Staneva Robert Rentzsch1. Skills-matching and skills intelligence through curated and data-driven ontologies. *Proceedings of the DELFI Workshops 2020, Heidelberg, Germany*, 2020.

[45] Stichting Samenwerking Beroepsonderwijs Bedrijfsleven (SBB). Qualifications and occupations. `https://www.s-bb.nl`, 2024. Accessed: 2024-12-27.

[46] Hans G. Schuetze and Maria Slowey. Participation and exclusion: A comparative analysis of non-traditional students and lifelong learners in higher education. *Higher Education*, 44(3):309–327, 2002.

[47] Ulrich Teichler. *Higher Education Systems: Conceptual Frameworks, Comparative Perspectives, Empirical Findings*. Springer, 2017.

[48] Suzan Verberne Thijmen Bijl, Niels van Weeren. Efficient course recommendations with t5-based ranking and summarization. *ReNeuIR 2024 (at SIGIR 2024) - 3rd Workshop on Reaching Efficiency in Neural Information Retrieval, 18 July, 2024, Washington D.C, USA*, 2024.

[49] UNESCO. *ISCED Fields of Education and Training 2013 (ISCED-F 2013): Manual to accompany the International Standard Classification of Education 2011*. UNESCO Institute for Statistics, 2014.

[50] UNESCO. Education for all 2000-2015: Achievements and challenges, 2015.

[51] UNESCO. Global convention on the recognition of qualifications concerning higher education. `https://unesdoc.unesco.org/ark:/48223/pf0000373115`, 2019.

[52] UNESCO. Reimagining our futures together: A new social contract for education. `https://unesdoc.unesco.org/ark:/48223/pf0000379707`, 2021.

[53] Qingheng Zhou, Yaliang Li, Nan Du Wang, Bolin Ding, and Reynold Cheng. Cross-lingual taxonomy alignment with bilingual knowledge graph embeddings. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 1615–1624, 2020.