NATURAL LANGUAGE PROCESSING IN LIFELONG LEARNING CHOICES - A CASE OF FINLAND

Lappeenranta–Lahti University of Technology LUT

Master’s Programme in Business Administration, Master's thesis

2023

Irina Zheleiko

Examiner(s): Associate Professor, Jan Stoklasa, PH. D.

Professor, Pasi Luukka, D.Sc. (Tech.)

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Business and Management

Business Administration

Irina Zheleiko

Natural language processing in lifelong learning choices - a case of Finland

Master’s thesis

2023

55 pages, 4 figures, 9 tables and 0 appendices

Examiner(s): Associate professor Jan Stoklasa, PH.D. and professor Pasi Luukka, D.Sc. (Tech.)

Keywords: Natural language processing, Sentence Transformers, educational courses, taxonomy-based information extraction.

Such techniques of Natural Language Processing as information extraction and semantic text labelling had been widely utilised in recruitment sphere to decrease the labour and time resources needed to analyse CVs or labour market’s trends. However, the application of such techniques and establishing link between demand for the workforce and education providing organizations is yet to be established. In the current thesis the ideas on processing educational courses descriptions texts is provided in attempt to facilitate the information exchange between the needs of the labour market and skills supply from the educational establishments.

In the literature review the analysis of the most recent methods in natural language processing methods is provided (Word2Vec, NER, Sentence Transformers) as well as commentary on their current implementations in labour market related spheres. In the empirical section state-of-the-art SBERT language model is applied to the collected open university courses’ descriptions in order to extract concrete skills from the and then the performance of the SBERT model is accessed through such metrics as precision, recall and f-score, yielding the F-score of 70.4%. As a result, an example of comparison between the skills supplies as identified by Finnish open universities educational courses and demand as identified by the job descriptions data is provided. In conclusion, research paper’s possible managerial applications and theoretical contribution are included.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my academic advisor, who was very patient with me and provided many useful insights needed to make this thesis comprehensive.

Abbreviations

AI - Artificial Intelligence

BERT - Bidirectional Encoder Representations from Transformers

IIBA - International Institute of Business Analysis

NER - Named-entity recognition

NLP – Natural Language Processing

RAKE - Rapid Automatic Keyword Extraction

SBERT - Sentence-BERT

ST – SentenceTransformers framework

Table of contents

List of figures

List of tables

List of formulas:

1 Introduction

Nowadays it is hard to come across a field that is not characterized by rapid technological development. Within all spheres from engineering to business new concepts and technologies are constantly emerging. This is especially true for the fast-paced Information system domain, where technological development spills into the labour market and its needs. Now, both IT specialists and the companies are in constant need to monitor the relevance of their data analytics and development teams's skillsets in order to be competitive on the market. The demand and interest for the continuous learning opportunities sparked offer of numerous educational courses platforms, such as MOOC.fi (MOOC.fi, 2022) developed by the university of Helsinki's Department of Computer Science, Skillshare and Coursera (Skillshare, 2022; Coursera, 2022) to name a few that are offering numerous educational programmes to people despite their current level in a variety of subjects. Moreover, institutions of higher education are constantly renewing their degree programs in hopes that their graduates are competitive on the labour market. Such wide range of opportunities creates the difficulty of choice of the needed program or course not only for the enrolees themselves but also for the leaders of IT-teams who may want to update certain skills of the team members. This thesis would be concerned with data-analytics and IT related competences, however, the approach is general and could potentially be applied to other occupations and fields as well, given that the taxonomy of competences for the occupation in questionsexists or will be developed.

1.1 Research gap

The recent years have seen rapid development of Natural Language Processing tools. Such fast development may be evident of how valuable it is to the companies' operations (Gruetzemacher R., Paradice D., 2022). The applications of the research progress had been found in many different fields from medical sphere (Stember J., Shalu H., 2021), where SBERT (NLP approach) was used in order to classify 3D MRI brain scans to finance (Chan C.S., Pethe C. and Skiena S., 2021), where BERT was used in order to predict the crowdfunding outcome of a project. The recruitment sphere has not been left out either.

Within it in the past few years the body of research concerning the potential applications of the language models had been growing at a quite high speed. These had led to creation of special services like Skills-Ovate and Zippia facilitating the in-demand competencies analysis. Many methods were developed concerning the need to match job positions and job seekers through starting from matching algorithms (Zhang Y., Yang C. and Niu Z., 2014) to the more recent trend of utilizing text-based matching (Zhu C., et al., 2018, Bian S., et al., 2019). Despite that, when it comes to the texts of educational programs, the annotating efforts had been addressed with quite limited attention. Filling this research gap and bringing more attention to this field of NLP-based methods applications is one of the purposes of the current study.

1.2 Relevance of the Study

Bringing more attention to the topic of educational courses annotation with skills that students may acquire opens a lot of possibilities. First it would allow for more fast and effective analysis of the potentially supplied skills, which in turn, would allow for comparison between them and the skills demanded by the employee. By establishing this rapidly accessed link between educational programs and labour market the educational institutions could adjust their curricula so that they are able to supply the labour market with specialists having relevant skillsets, while job seekers could get a better understanding of their career opportunities based on the education they may be applying for. Overall, the topic of educational courses annotation with skills seems to be promising and relevant for both business and educational sector.

1.3 Research questions and assumptions

The objective of the current thesis is to come up with a way to extract skills from educational courses descriptions and use those as a basis for comparison the supply of the competences to the demand of the labour market. Therefore, the main research questions are:

- Can the text used in the educational program’s description be potentially used for skills/competencies extraction through the usage of a Sentence Transformer model?

- Is it possible to apply NLP methods in order to assess how well the market's need in skills is satisfied by the existing courses and identify skills that are not yet covered enough and could be potentially added to the study programmes?

It is worth mentioning, that his research operates on several assumptions.

- The first one being that when choosing a course, the potential attendee has the access to course description and bases his choice on the provided information, therefore it is assumed that the educational course's description depicts the contents of the course explicitly. And is sufficient to infer what skills the attendee will acquire as a result of the course completion.
- In the recent studies the idea of the need to introduce unified vocabulary and writing syntax for educational courses descriptions had been gathering more popularity (ex. Wunderlich J. and Tilebein M., 2019). However, it is still to be achieved. Therefore, due to the absence of an uniform way of educational course descriptions, and generally analysed offered courses having the same amount of credits, it is also assumed that courses that contain less mentioned skills offer a more detailed view on them, while courses offering several – provide a more general overview on the mentioned concepts.
- Due to the limited resources when writing a thesis, the 'Golden corpus' offered as a possible method of accessing the model performance was created based on own general judgment and meticulous analysis of existing competency models and the utilized taxonomy. And it is assumed that this assessment was done in the correct manner.

1.4 Goal

The main goal of the current research paper is to propose a possible solution to establishing a link between educational courses supply and labour market demands for skills through utilizing a unifying taxonomy. And to outline a possible recommendation algorithm that is able to suggest the most relevant educational courses needed to fill a certain knowledge gap of an employee/team member. In order to achieve the set goal, a taxonomy of IT-related skills, is going to be utilized, and natural language methods will be applied. Therefore, the following tasks need to be completed:

- Analyse existing findings in the examined field (literature review)
- Collect and aggregate the data on educational programmes
- Build & test the model that extracts competencies and skills from the educational programmes' description texts
- Propose a method for matching elicited skills and educational courses

1.5 Structure of the thesis

Current thesis is organized in six parts.

In the first chapter the justification behind the thesis is provided, which consists of the current situation assessment, main goal identification, list of tasks needed in order to achieve the proposed goals and technical requirements needed to evaluate whether the goal had been reached.

Within the second chapter one may find description of the techniques' companies use to evaluate the skillset of employees, and what services infer labour market insights based on the job advertisement. Moreover, overview of the recent most popular methods in NLP is provided.

In the third chapter solution development is provided with detailed description of the applied model. Results and possible validation is provided. What is more, comparison of the in-demand competencies (as derived from previous research from job description data analysis) and those offered by the educational courses is offered.

The fourth chapter expands on potential use cases and managerial applications of the proposed algorithms and its results.

In the fifth chapter limitations of the current research are provided so as ideas for future research.

Final, sixth section is the conclusion with description of main steps taken and the achieved results.

2 Overview of the current situation of competences handling-related methods

In the current chapter techniques of assessing employees’ competencies used in the companies are described, followed by frameworks on which they rely and which served as the basis of the development of the utilized taxonomy and other international services relying on the classification systems. In the last part of the section an overview of the most popular NLP methods is provided.

2.1 Overview of existing employees’ evaluation techniques

In the rapid changes of technical process more and more companies practicing digitalization and automation process are marking employees and their skillset as the most crucial component in their process (Tortorella, G., et al, 2018). The deficiency of core competences within the employee may lead to negative organizational performance (Yasar, M.F., Ünal, Ö.F., Zaim, H., 2013). Hence for companies the assessment of the employees’ skillsets is a field of interest. Due to complexity of the task, as of now, there is no one concrete tool, but rather an availability of several methods to choose from. After the assessment of the literature review, the current division into groups may be offered:

1. Test

   When it comes to hard-skills and quantitative skills, these can be assessed through tests that the employee should complete in order to determine his proficiency in a set of competencies in question. The drawback of the current method is the fact that it may not be applicable to individual’s soft skills, which are more complex and personal ones (Efremova N., Shapovalova O., Huseynova A., 2020).

2. Self-Assessment

   The individual in question is asked to rate their competencies on a scale, stating in which he feels himself sufficient, proficient or lacking etc. The drawback of this approach is the bias that people may possess when it comes to their self-evaluation. In fact, less competent people are more likely to rate their capabilities higher than their more able counterparts (Holt, J., Perry S., 2011).

3. Assessment by others

   3.1. Assessment by the employer

   The assessment of employees' competencies is done by the employer or manager himself, could be performed through completing a questionary, where certain competencies can be assessed through a scale, for example from 1 to 5 (Vukajlović D., Brzaković2 M., 2016).

   3.2. Assessment by the colleagues/team members

   Here the employer may ask team members to offer their assessments of a certain employee: what are his strongest and weakest points. The benefit associated with this type of assessment is the fact, that it is more of an objective approach, since here several individuals offer their insights, rather than just one person (manager), who due to human nature could be potentially biased (Business Finance Articles, 2021).

   3.3. Feedback from clients

   If applicable (if clients have a closer contact to the employees), they can be asked to provide feedback on their experiences with a certain employee. For example, through NPS, or Net Promoter Score, where clients are asked to rate on a scale from 1 to 10, how likely they are to recommend, or promote services/products that the company had been delivering to them. NPS then is calculated as difference between the percentage of promoting customers (during rating provided 9 or 10) and 'detractors' - scored 0 to 6. However, here, an aggregated assessment is offered, without seeing the concrete competences of the team.

4. Accomplishments

   For a more technical approach, one can use a set of metrics in order to evaluate an employee's skillset. For example, according to Sales Enablement Analytics Report 2019 (Sales Enablement, 2019) many sales enablement professional relies on activity-based metrics such as the number of training sessions delivered, number of sales personnel achieving quotas etc. Performance metrics, such as time needed to complete a project, times the tasks had to be re-done, etc. are widely used as well (Factorial HR, 2022).

5. A 360-Degree Approach

   Is an approach that merges both self-assessment and the assessment by the third parties as well. Here, not only the employee provides his self-assessment but also do his teammates, line colleagues, and even clients and outside agents and vendors. It usually proceeds as follows: the person himself and about 8 to 12 people (Custom Insight, 2022) receive an anonymously completed feedback forms, asking to rate the individual in question on a variety of work-related competencies on a scale. The statistics and insights from the forms are gathered and through them an analysis of individual's competencies is provided, including his strength, weaknesses and point of further growth.
6. Aggregational methods

   Some of the companies offering the assessment of the employee's or teams' skillsets and competencies (eg. Digital City Planner, which taxonomy had been utilized for the purposes of the current thesis) use the combination off of the techniques above in order to access the capabilities of the employee or the whole team to the highest possible degree.

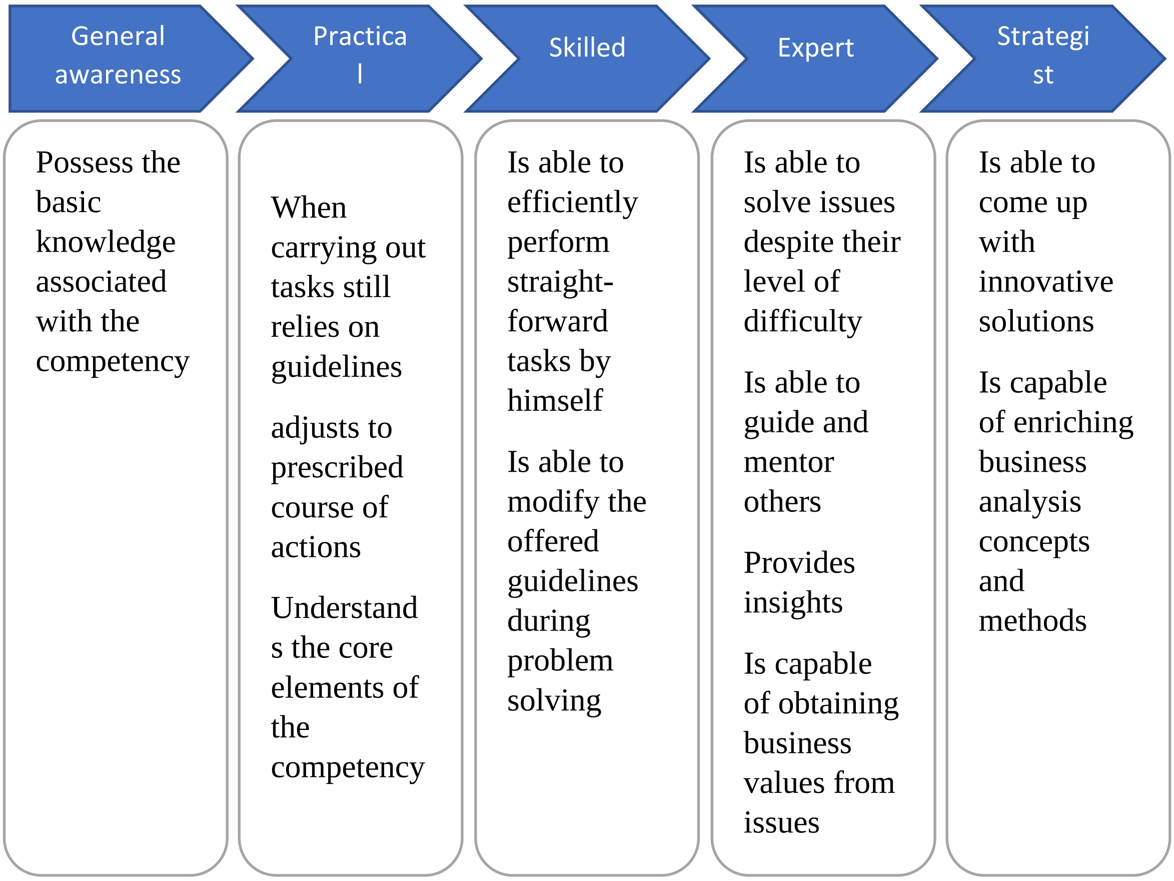
Both self- and assessment by the third party rely on the availability of the competencies models: lists of competencies that are being rated by the reviewers. Such models contain and describe a collection of skills, personal qualities and knowledge that are required to be sufficient in performing your duties as an employee. These models are widely used by HR during the process of choosing candidates, providing training to newcomers, monitoring effectiveness of current employees and so on. Hard skills can be quite tangible and can be evaluated more easily (for example, how proficient is individual in Python, what packages he may effectively use to complete what type of tasks). Some knowledge can be measured as well (how many languages does the candidate know). However, there are more complex ones, such as "critical thinking' or 'levels of desire to learn and self-develop', evaluation of which becomes more intricate and difficult. In order to able to achieve this task, set of interviews, questionaries and focus groups can be organized in order to find the most effective strategy of dealing with these un-tangible competencies (Seema S., 2016).

2.2 Overview of the existing competency models and frameworks

When it comes to IT related competencies, in order to assess employee's or team's skillset one may rely on several available competency models, which in detail outline the competencies and their potential levels of expertise. required for successful completion of IT and data analysis related tasks.

2.2.1 IIBA Business Analysis Competency Model

This model was developed by the International Institute of Business Analysis (International Institute of Business Analysis, 2017) in order to qualify competences that are expected from an exemplary business analyst. These models employs the scale ranging from 1 to 5, and corresponding proficiency levels:



General awareness	Practical	Skilled	Expert	Strategist
Possess the basic knowledge associated with the competency	When carrying out tasks still relies on guidelines adjusts to prescribed course of actions Understands the core elements of the competency	Is able to efficiently perform straight-forward tasks by himself Is able to modify the offered guidelines during problem solving	Is able to solve issues despite their level of difficulty Is able to guide and mentor others Provides insights Is capable of obtaining business values from issues	Is able to come up with innovative solutions Is capable of enriching business analysis concepts and methods

Figure 1. Competency levels as outlined in the IIBA Business Analysis Competency Model (International Institute of Business Analysis, 2017).

The defined areas of knowledge are:

- Solution Evaluation

  The ability to evaluate the company’s performance and suggest ways to eliminate obstacles affecting the potential value created by the company
- Business Analysis Planning and Monitoring

  The ability of the specialist to carry out and manage the workflow of the fellow business analysts and stakeholders.
- Requirements Analysis and Design Definition

  The ability of the business analyst to carry out an elicitation actions and support the acquired results, which are later communicated to the stakeholders. What is more, this knowledge area also includes the continuous cooperation between the business analysts and the stakeholders during all business analysis activities.
- Requirements Life Cycle Management

  The ability of business analysts to manage requirements and design information through its whole life cycle
- Strategy Analysis

  The ability of business analyst to elicit business needs that may be of strategical importance. After which he or she should also be able to tackle the identified need and coordinate the needed course of action with other strategies despite of their hierarchy level
- Elicitation and Collaboration

  The ability of the business analyst to map out requirements educed as the result of the elicitation activities. Outline requirements. Justify and confirm the acquired information. Educe possible solutions and evaluate the value that the business could acquire from the set of educed potential solutions.

The competency model also identifies underlying competences business analysts needs to possess in order to be efficient in his or her tasks. These include their personal characteristics and qualities, knowledge and their approach of carrying out tasks. These include: analytical thinking and problem solving, business knowledge (of the company, market, industry, recent business trends etc.), interaction skills and communication, enabling them to perform effectively within a team, collaborate with stakeholders and deliver their findings in a clear fashion; behavioral characteristics such as adaptability, accountability etc.; and the last, but not the least: tools and technology (a range of software applications that could be used in the business analytics purposes as well as communication and improving gross productiveness).

2.2.2 TOGAF Architecture Skills Framework

This framework of skills was developed by The Open Group (Open Group, 2006) to characterize the capabilities of the specialists working in the area of enterprise architecture. It was developed to describe skillset of a typical IT architecture team, rather than just one specialist.

The skillset of the efficient team consists of seven main categories:

- Generic Skills

  What can be described as soft skills, these include leadership, ability to work within a team, effective communication etc.
- Business Skills and Methods

  The ability to perform in business cases, work with business processes, perform strategic planning, mange budgeting etc.
- Enterprise Architecture Skills

  These usually include design of business processes, data, role, organization, etc., solution modelling, benefits analysis and so on.
- Program or Project Management Skills

  These skills are often comprised of change, project value and program management.
- IT General Knowledge Skills

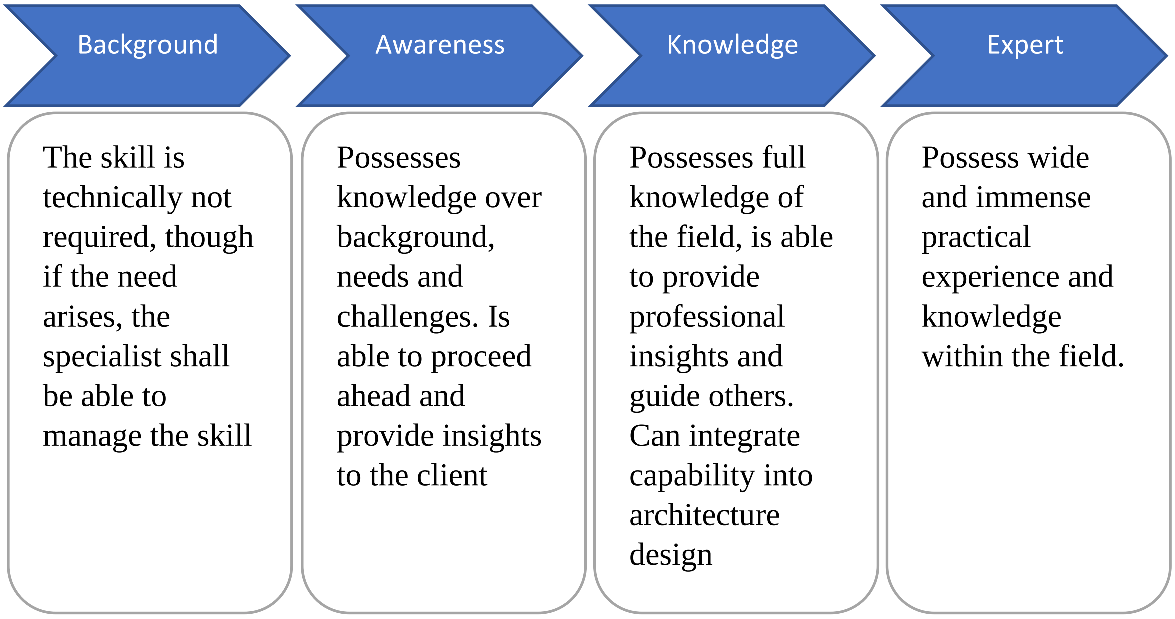
  Within these skills are usually found knowledge of programming languages, brokering applications, storage management, networks etc.
- Technical IT Skills

  These usually are comprised by software engineering, User interface, data management and so on.

- Legal Environment

  Here it is implied that the team possess knowledge over contract law, data protection laws, fraud related issues and associated course of actions and others

Unalike IIBA Business Analysis Competency Model, TOGAF Architecture skills framework rates specialists' proficiency levels on the four-level scale:



Background

The skill is technically not required, though if the need arises, the specialist shall be able to manage the skill

Awareness

Possesses knowledge over background, needs and challenges. Is able to proceed ahead and provide insights to the client

Knowledge

Possesses full knowledge of the field, is able to provide professional insights and guide others. Can integrate capability into architecture design

Expert

Possess wide and immense practical experience and knowledge within the field.

Figure 2. Proficiency levels within a competence as identified in the TOGAF Architecture Skill Framework (Open Group, 2006).

2.2.3 FEAPO

The main goal of the FEAPO (FEAPO, 2018) competency model is distinguishing between different architecture roles, specifically tailored to be applied to the team. It follows the similar structure of TOGAF which is organized as shown in the table 1:

	Role 1	Role 2	Role 3
Competency Group			
Competency	Level	Level	Level
Competency	Level	Level	Level
Competency Group			

Competency	Level	Level	Level

Table 1. FEAPO competency model structure

Competency group is an umbrella term for all the associated sub-competencies that follow in this category.

Competency is the ability that specialist might possess to perform tasks successfully. These can be learnt from a range of sources: formal education, previous work experience, personal experiences etc.

Role is a set of competencies that FEAPO specialists marked as sufficient and important during their work experiences. It is worth mentioning that the roles are not assigned to people individually, nor they are set in stone. Some of the competencies may overlap between specialists and some individuals may perform several roles if required.

In this competency model the proficiency level is determined on a 1 to 7 scale. Where proficiency levels are the following:
“Follow -> Assist -> Apply -> Enable -> Ensure and Advise -> Initiate and influence -> Set strategy, inspire and mobilize”

2.2.4 Existing services relying on job description data for labour market’s analysis

In order to develop these frameworks, Business and state bodies had been creating tools so as to facilitate statistical analysis of the job titles and their competencies starting all the way back to 1960s (Rentzsch R., Staneva M., 2020). One of the most prominent and well-known classification of this nature is “The international Standard classification of Occupation”, or ISCO for short. With the surge of differentiation within the job market during 1990s first competencies taxonomies had been developed, such as O*NET, a database of Occupational Information Network, developed in the US . When it comes to Europe in the past 20 years classification systems, such as ESCO (European Skills, Competencies, Qualifications and Occupations) had been developed. Both ESCO and O*NET are taxonomies, exploring links between concepts, competencies and other ontologies.

On the basis of these taxonomies, through the usage of recent information technologies and job descriptions' analysis as proxy for labour market demand several services had been developed. An example of this would be Skills-OVATE (Cedefop, 2022) for European market and Zippia (Zippia, 2022) for the American market.

Skills-OVATE

This service analyses demands and needs expressed by employers through the millions of job descriptions aggregated from 28 European countries and a variety of sources through the joint forces of Cedefop and Eurostat (Cedefop, 2022). One of the strongest benefits that the service prides itself on - is how up-to-date their information is. The insights provided by the service is in the form of dashboards that the user may interact with: through the dashboard the user can filter regions of countries, industry sectors and job positions. Additional function at Skills-OVATE's disposal is provided analytics on the in-demand skills with country, field and general labour-related trends.

Skills-OVATE utilizes two skill classification. The aforementioned ESCO and O*NET. The former is comprised of 3 levels. The first one is divided in four groups: skills, knowledge. Attitudes and values and languages. On the second level is split into spheres of application, such as "art and humanities", "business, administration law" etc. On the final, third level, one may find the names of the skills themselves, such as "verbal communication" or "project management" etc.

O*NET classification is comprised of several levels as well. On the first one we can find such groups as: knowledge, work activities, work styles, skills, technology skills and tools and abilities. This allows for a more detailed view, since technical and more general, akin to soft skills are divided into different subgroups. The second level is similar to that of the ESCO classification in a sense that it is also comprised from fields of identified skills applications. On the final level, concrete skill names are given, and their variety is greater compared to that of the ESCO classification.

When it comes to information extraction, Cedefop relies on such ML algorithms as sentiment analysis, n-gram and NER (Cedefop, 2021).

Zippia

Zippia (Zippia, 2022) is another example of a service using job postings as a proxy for labour market trends. It aggretes numerous job descriptions as well (including requirements for the potential employee, salary and companies from which the job vacancies had been collected) to offer analytical insights on the expected salaries based on working experience, region of the job vacancy, industry and the education lever of the interest.

When it comes to skills, Zippia is able to provide skills output with percentage of their occurrence in resumes vice versa it can show, what vacancies are requiring a skill of interest. However, in this case the service does not utilize any taxonomy as a basis for it skills extractions, but rather simply uses keyword extraction techniques. As a result, the elicited skills are too vague.

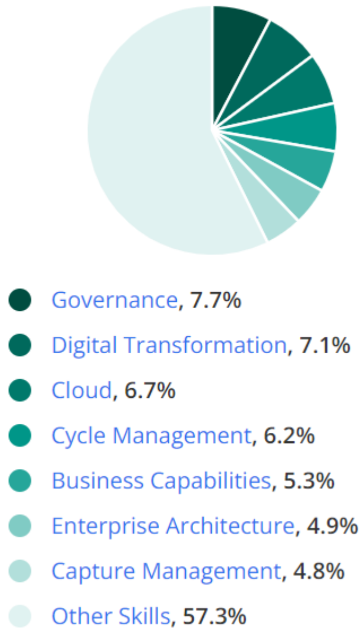


Figure 3. Eight most common skills based on Business Architect Resumes in 2022 (Zippia, 2022).

In the figure 3 you can see an example of skills analysis for Business Architect as identified by Zippia. Such skills as “Governance” and “Cloud” are not very informative and just by looking at those no concrete understanding of what exactly the identified skills require the applicant to be able to do can be inferred from the output.

2.3 Utilized taxonomy description

Taxonomy is essentially a formally expressed structure of objects' classes within a certain domain. Taxonomy possess a certain hierarchical structure with parent-child relations between the concepts and a controlled vocabulary (Otendo, 2022). Skills taxonomy is ordered list of skills needed for successful completion of tasks within a company (Linkedin, 2022).

For the purposes of this paper a taxonomy of skills developed by Digital City Planner (Digital City Planner, 2022) was utilized. Digital City Planner is a consulting Helsinki-based start-up that offers quite a broad spectrum of knowledge-management related services and products, as well as enterprise architecture areas. During its more than 15 years of experience it had provided more than 150 courses, published more than 100 research works and, what is mostly important, successfully accomplished more than 50 projects. In order to do so the company has developed their own taxonomy for which they had collected roles and competencies from all the aforementioned competencies models and taxonomies, while consulting with experts in the field. It utilizes a 1 to 5 point scale like the IIBA competency model, and is organized alike TOGAF and FEAPO models to facilitate its usage for the team assessment. What is more, in this taxonomy skills for a broader specter of IT-related specialists (including Data analysts, business analysts, enterprise architects, System engineers etc.). In total, the taxonomy includes 399 competencies.

Competence is used to refer to a set of employee's ability to perform their tasks and includes knowledge, hard and soft skills. In order to provide a more detailed analysis of the educational courses description texts' assessment additional categories were introduced: products, tools and frameworks. Therefore, the utilized taxonomy operates on the list of types of competencies as described below.

Knowledge: expertise in a certain knowledge area (eg. Knowledge of the company's current business process)

Hard skills relate to more technical skills needed to perform job related tasks (data analytics, budgeting and accounting etc.)

Soft Skills related to personal characteristics of the employee and represent his or her modus operandi – their ways in which they approach tasks etc. They also include interpersonal skills, communication, team working abilities etc.

Product: Ability to utilize a certain product/program (eg. MS Word)
Framework: Ability to utilize a framework – (eg. Scrum)
Tool: Ability to utilize a tool needed to perform a technical task (eg. Python)
Database: Ability to work with a certain database/database type (eg. NoSQL)

2.4 Overview of the education market in Finland

One of the most prominent specifications of the Finnish education market is the wide range of open university courses available. These are offered by every of 14 Finnish universities (Opintopolku, 2022; Education Finland, 2017). These open university's courses are available to anyone independent of their background, prior education, or age. What is more, these courses are compatible with degree programs of other university, insuring sufficient levels of materials' deepness, compatible with both master's and bachelor's degree level. This allows a person upon completing such course to acquire a high level of subject's understanding, while choosing only the subject of interest, without ither courses that might be needed in order to get a certain degree. This is a short-term option to enrich your knowledge of a certain field in concise time period.

2.5 Overview of the main Natural Language processing concepts

In the past several decades, more and more industries had been relying on the automation of tasks and usage of AI and21st century had been marked by the the 4th industrial revolution (IBM. 2022), which was characterized by rapid movement towards automation of manufacturing processes, utilizing newly developed technologies. However, the influence of Industry 4.0 was not limited to industries and manufacturing operations. Human resources management was influenced at a great degree by it as well (Bayraktar O., Ataç C., 2018).

It is believed that labour market structures are going to be changed to a significant degree due to trends in the technological breakthroughs. The whole core of in-demand competencies and skills needed for successful tasks running is likely to change rapidly with new skills and knowledge areas showing up (World Economic Forum, 2018). In order to be able to keep up with ever-changing trends within the job markets many researchers proposed the usage of AI to automate HR-related processes (Djumalieva J., Sleeman C., 2018; Dadzie A., et al., 2017). When it comes to the recruitment as of now a lot of processes are done manually, which requires a lot of time and computational resources (Mhamdi D., et al., 2020). As a potential answer to automation of the recruitment process and transformation of the practices NLP was considered to be "the most promising technology" (Manatal, 2022).

NLP, or Natural Language processing is a field within the machine learning that covers a wide range of topics related to understanding, computing and assessment of human languages (Otter D., Medina J. and Kalita K., 2019). The eight core NLP labelling tasks are: part-of-speech tagging (syntactic task of words labeling as nouns, verbs etc.), constituent labelling (like the previous task, but here the phrases are identified as noun phrase, adjective phrase etc.), dependency labelling (the dependencies between words in the phrases are labeled), named entity labeling (NER task of assigning categories to words: organization, person etc.), semantic role labelling (identifying predicate-argument structure of the phrase, based on the semantic meaning, for example what word relates to subject and which to object), semantic proto-role (are more complicated version of the semantic role labelling, here the more detailed semantic attributes are labelled, such as state of awareness for example), relation classification (task of understanding the "real word relation" that can be found between the two identified entities) and coreference. (Tenney I., et al., 2019). For the purpose of this study, the main task is the coreference, which is responsible for finding out whether two tokens mean the same entity.

2.5.1 Text labeling as a tool for information extraction

During the past years information extraction has been becoming more intricate with the introduction of data mining, new programming and business analytics tools (Kamran M., Anjum M., 2017). One of the methods with a growing body of research was ontology-based information extraction. Through the reliance on semantic matching the authors were able to efficiently extract information using this approach (Mestrovi A., Cal. A., 2016; Ramli F., Noah S., Kurniawan T., 2016). This approach had been especially popular in the field of research related to skills and recruiting practices. Here the researchers had utilized taxonomy-based information extraction approaches in order to successfully extract skills from the online job adverts (Djumalieva J., Sleeman C., 2018, Sibarani et al., 2019). In the work of Rentzsch R., Staneva M., (2020) the authors outline many spheres where skills-related taxonomies and ontologies, powered through the usage of NLP and ML algorithms are applied:

- Machine-readable annotation:
  Here with the help of ontologies commercial services help employee seeking companies create job descriptions to be posted on recruiting platforms. Some of the services, such as ChatGPT (OpenAI, 2022) - an enormous language model developed by OpenAi is able to generate cover letters, based on the input provided in the form of a CV.

- Semantic searching
  Allows for a more detailed searching abilities, since unalike keyword-based searches, semantic-based ones are able to capture intricate relationships between concepts that are provided in the taxonomies, therefore allowing for taking the account the semantic meaning (Search Engine Journal, 2022). The most simple example of it is the machine's ability to analyze synonyms and treat them as the same entity, when semantic meaning of a word/phrase is considered.

- Semantic matching
  This can be seen as the logical continuation of semantic searching. Here the whole documents may be matched between each other: such as CVs and job description data. This a concept that had been very popular in the research field and successful implementation of such technique can be seen in the works of Fazel-Zarandi, M., & Fox, M.S. (2009), Grüger, Joscha & Dr. Schneider, Georg. (2019) and Gugnani, Akshay & Misra, Hemant. (2020).
- Multilingualism
  Classification systems like DISCO and ESCO, available in several languages allow for annotation of both vacancies and resumes in a variety of languages, which enhances information circulation between both employers and job applicants around the globe.

- Labour and market analysis

  This field of taxonomies (e.g. ESCO, 2022) and NLP applications allows for monitoring recruiting platforms, forums etc. and, as result, provides information of current trends on the labour market, in-demand skills and can potentially help balance out workforce supply and demand of skills and competencies. A platforms relying on this field of research such as Skills OVATE had been mentioned in the above section.

However, Rentzsch R. and Staneva M. (2020) also pay attention that despite the popularity of annotating job advertisements, CVs and resumes and matching those, very little attention had been given to annotation of educational programs and courses with skills that they could potentially deliver in both private- and public-based research.

2.5.2 Word2Vec

In order for an NLP algorithm to become applicable to the text it first needs to be encoded, or embedded, turning a string into a continuous numeric representations. The simplest way to do so is to implement a Word2Vec model. These models upon being fed big corpus of text data create vector space with high dimension order where each unique word is assigned to each of the vectors on this vector space. Despite the model's relative simplicity and popularity, there are two main issues that arise when using such a model:

- When met with a new word that the model had not encountered during its training, it will assign randomly generated vector to it, meaning that the new word will not be associated with any other words in a sensible way. Therefore, meaning that the model cannot perform well on new set of data that it had not seen before (Kandi M., 2018).

- The word2vec model encodes singular words and thus may not capture the semantic meaning of a phrase as a whole (Sibarani et al., 2019).

2.5.3 NER

NER, or named entity recognition is a NLP task which main goal is to elicit entities from an annotated text and classify them into categories: such as person, location, time etc. NER is typically presented in several configurations.

The first, the most basic type of NER is dictionary-based. This method relies on the existence of a dictionary with labeled entities. When applied to text it simply matches entities found in the texts with those that are contained in the NER's vocabulary.

The second method is rule-based. These rules are defined beforehand and are used on order to determine the IE method. These could be presented in the form of pattern-based rules, where morphological patterns are specified. For example, if a noun follows after a verb with a preposition – it is most likely to be a location.

The drawback of both aforementioned approaches is the fact that they rely on the previously existing manually developed lists of entities or set of rules and patterns, which take time and resources to create and may not be updated as often as they should in order to be able to work with new data.

Recently there had been developed a third approach that relies on machine learning techniques that can be either supervised or unsupervised, enabling the user to a pre-trained models that do not require to develop a specially tailored dictionary or set of rules. An example of such a toolkit is GATE (Gate, 2022) developed by the University of Sheffield. It is worth mentioning that some of the Python packages, for instance, SpaCy and NLTK have own NER systems ingrained in their pipelines.

2.5.4 RAKE

RAKE, or Rapid Automatic Keyword Extraction is an unsupervised method used in order to extract keywords from an individual document, independent of domain and language (Rose S., Engel D., Cramer N. and Cowley W., 2010). In the sentence not all the words have the same level of meaningfulness, despite stop words, that don't provide significant lexical value (functional words: articles, joinders, etc.), there are less likely to be important when getting the meaning of the sentence. Embedding of a sentence takes a lot of operational memory (Rose S. et al, 2010), therefore it would be beneficial to get rid of the noise-introducing words and only keep those that would offer concise representation of the text's contents. According to Rose S. and other developers of the RAKE, this method offers an "extremely efficient" solution for the keywords extracting task, whilst being applicable to a wide range of domains, including new ones, despite the type of them, and what is more, highly productive on texts, that were not developed according to "specific grammar conventions".

The working algorithm of RAKE is organized as follows:

1) As an input, the model is offered a list of the stop words, sets of phrase and word delimiters, dividing a phrase into "candidate keywords", or, in other words, core-expressing chunks.
2) Then co-occurrences of words found in these chunks are used for identification of words associations that are specific for the analyzed document in question.
3) Scoring of the candidate keywords which is calculated as a sum of the member word scores contained within each individual candidate. The evaluated scores are: word frequency (freq(w)), word degree (deg(w)) and ratio of degree to frequency (deg(w)/freq(w)).
4) After the computation for all keywords had been carried out, the top T candidates with the highest scores are selected. The authors in accordance, with the work by Mihalcea and Tarau (2004) use the number equal to one third of the identified co-occurrence found in the step 2 as the T, to choose the final amount of the extracted keywords.

It is also worth noting that RAKE does not disregard keywords that do contain a stop word between them if it is integral to its meaning. For example, if the combination "master degree in data science" were to occur at least two times within same document, and in the same order, the whole phrase would be seen as a keyword, so despite the presence of the stop word "in", the information is not lost.

The advantages RAKE holds over other keyword extractors, such as TextRank and Ngram method was described in the work of (Rose S. H, Automatic keyword extraction from individual documents). The authors had compared most popular keyword extractors and found, that RAKE is more efficient in terms of estimation compared to TextRank, while simultaneously reaching larger recall value with no loss in the recall metric. To a similar conclusion came Rajashekharaiah K. M., Ganiger S. (2018) finding that RAKE outperforms both TF-IDF and TextRank significantly when it comes to recall, while having no significant difference in the F-score.

2.5.5 BERT, SBERT family of models

Bert, or Bidirectional Encoder Representations from Transformers is a language representation model that is able to achieve state-of-the-art performance (Delvin J., Chang M., 2019; Ptiček M., 2021) in a wide range of NLP-related tasks, including text classification (Sun et al. 2019). BERT-based models were found to outperform GPT (by nearly 2 F1 points on average), ELMo (by nearly 2.7 F1 points, while offering an almost 20% decrease in the error term) in the work of Tennet I, et al., 2019). To similar conclusions of BERT-type of model superiority came Ptiček M. (2021) and Delvin et al., (2018).

SBERT, or a Siamese BERT relies on the usage of Siamese (Towards Data Science, 2022) and triplet network structures. Siamese Neural Networks (SNN) were developed in order to enhance the amount of resources both time and computationally wise required for comparing items. This is achieved through subjecting two inputs to identical subnetworks. After to the embedded objects a loss function is applied. This loss function's task is to minimise the distance between similar items, while maximising the distance between different items. Due to the subnetworks being the same, lesser amount of computational power is required in order to run SBERT model. In fact, Reimers and Gurevych (2019) argue that SBERT reduces the time needed to discover the most alike pairs of sentences from 65 hours to 5 seconds compared to just BERT's performance, while keeping comparable accuracy. All of this allows SBERT to capture the semantic value of the sentence as a whole and compare two sentences using the cosine-similarity score, while retaining each's semantic integrity.

When it comes to the usability of SBERT models several authors had characterised SBERT as the most tailored method to the task of comparison of semantic values as well as similarities for the sentences and single words as well (Reimers and Gurevych (2019); Bondielli A. and F. Marcelloni (2021)).

Below you can see the illustration of the SBERT working algorithm. At first both sentences are subjected to BERT network, which is responsible for the initial encoding of the sentence. Reimers and Gurevych then propose additional pooling layer that obtains a fixed sized sentence embedding. On this layer the output of BERT is fine-tuned through the usage of Siamese neural networks, which adjust the weights using the mean-squared error objective function, ensuring that the created embeddings contain semantic value that can be evaluated with cosine-similarity. U and V on the figure stand for the calculated sentence embeddings.

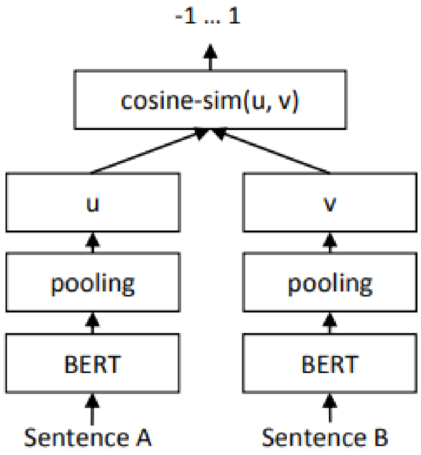


Figure 4. SBERT architecture at inference (Reimers and Gurevych, 2019).

The creators of the SBERT propose that for calculation of similarity either cosine similarity or Manhatten / Euclidean distance. For the purpose of this study cosine similarity was chosen.

2.5.6 Cosine Similarity

Cosine similarity is used to evaluate similarity between vectors. It is widely used in a variety of NLP-related algorithms. One of the typical usages of it is document similarity in particular (Towards Data Science, 2020). It operates by calculating the cosine of the angles between vectors; therefore, it takes into account the direction of the vectors (Han J., Kamber M. and Pei J., 2012). This direction has a very important influence over the semantic meaning (Predictive Hacks, 2022). For example if we wanted to find to a phrase “will be able to solve problems in an inventive manner” the most similar skills: “Applying creative thinking problem solving” or “Resolving technical problems”. The closest match is “Applying creative thinking problem solving” because in both of them creative/inventive approach to problem is implied. Despite the words not being the same, due to the direction of them being similar, they would be identified as having similar concepts. Eucludian distance on the other hand, does not take into the account the direction of the vectors, therefore it could provided us with a different match, judging by just the distance term (Machine Learning Mastery, 2020).

$$sim(A,B) = \frac{A*B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Formula 1. Calculation of Cosine similarity between two vectors, where A and B are vectors, Ai and Bi denote coordinates belonging to the vectors A and B respectfully.

$$d(A,B) = \sqrt{\sum_{i=1}^{n} (B_i - A_i)^2}$$

Formula 2. Calculation of Euclidean Distance between vectors, where A and B are vectors, Ai and Bi denote coordinates belonging to the vectors A and B respectfully.

3 Solution Development

In our previous work (Riabchenko A., Zheleiko I., 2022), 'Competency demand analysis based on job advertisement data', me and my colleague had tried several specifications and types of text analysis models in order to determine the most efficient one in terms of the F-score/F-measure. The tested models included NER (named entity recognition), SpaCy's inbuild similarity tool, custom trained on job description data Word2Vec model and averaging the phrase vector and SentenceTransformer family-type model. The models were applied to both sentences as a whole and sentences that had firstly been divided into key-phrases through the usage of RAKE algorithm. The parameter being changed within all the models was the threshold value, needed as a minimum to assign a label from the utilized taxonomy to an extracted skill/competence or mark it as a new competence, unknown to the current taxonomy. When comparing the models' efficiency, the best-performing one based on the F-score, was the SentenceTransformer model that had been subjected to the RAKE partitioning into the phrases. This model specification's results were far superior to both Spacy's similarity tool and the custom trained Word2Vec model. Therefore, for this research paper, the same type of model will be utilised, trying out several specifications of the threshold value in order to achieve the most efficient configuration.

3.1 Data description

For the purposes of the current thesis the data for the algorithm was collected from 13 universities (University of Lapland was not providing relevant subjects of courses in English as of data collection period time). The courses descriptions were collected from the section of open university programmes. The majority of the educational courses were collected from the StudyInfo (StudyInfo, 2022) and Opintopolku.fi (Opintopolku, 2022) web pages. Individual university own web pages were also utilised for data gathering.

Considering the speciality of the taxonomy, all skills related to programming, management, business administration, economics, corporate finance etc. were collected. Programmes' descriptions were collected regardless of the period when they should be completed, but usually the description texts are available only for those courses that are yet to start, and the application period is still ongoing (hence the collected courses were mainly for autumn 2022 semester with some for spring 2023 semester). Only those programmes were used which text description is available in English, due to availability of pre-trained in English machine learning algorithms and libraries.

In total 346 education courses descriptions were collected. From the whole text regarding the program, learning outcomes and course description sections were used. Each observation in the table contains city, where the university is based, name of the university and the course, text description and the fee of the course. The range of the fees is from 0 (some of the courses are free) up to105 euros. Sometimes there is a certain set fee needed in order to participate in the course, in other courses the fee is dependent on the amount of credits needed/completed. For unification purposes the maximum fee was calculated (the amount of all possible to acquire credits multiplied by the price of one credit). However, not for all the courses the information on the fee is available, therefore some had NaN values, hence the price was not used in the final ranking process, but one can see the price associated with the course when the information about the educational program is retrieved.

3.2 Data preprocessing

The data was processed in several steps. First of all, all the characters within the texts were transformed in the lower case, lists of bullet points were transformed into separate sentences. So that it would be easier for the algorithm to consume the text in the form of the ready to be analyzed sentences and lower the computational power needed. After that all the texts had been transformed into NLP objects (set of tokens) through the usage of en_core_web_sm, a spaCy's trained on written web text pipeline for English language, a small and lean model, optimized for CPU (SpaCy, 2022). Despite being almost 47 times smaller compared to en_core_web_lg model, that had been trained on larger corpus of data, en_core_web_sm F-scores when compared are either not statistically different from that of the larger model (Berragan C., Singleton A., Calafiore A. and Morley J., 2022) or even higher (Panoutsopoulos H., Brewster C., 2022).

Following the previous step, we then extract key-phrases through the usage of RAKE. Duplicates, if found within the set of keywords are deleted. In accordance with methodology described by Rose S., Engel D., etc (2010) and Mihalcea R and Tarau P (2004) for each sentence only one the number equal to one third of all calculated words co-occurrences are used as the final number of needed keywords. As was mentioned before, when extracting "keywords", RAKE does not only extract singular words, but phrases, if they are consistent and meaningful as well. Length of extracted phrases may vary from six to one word. These single words-phrases were kept, as sometimes skills can be represented by one word, such as names of tools (ability to use a tool for excecuting a certain task, for example a competency in a programming language, such as python or SQL), products (ability to utilise a particular products, such as a certain software or a program – Ms Excel), or database (e.g., "GreenPlum") (Sibarani, Scerri, Morales, etc., 2017). What is more, one-word skills may also include proficiency in the application of a certain frameworks, such as "Scrum", for instance. Hence, these one-word phrases were kept for each of the educational program text.

3.3 Sentence Transformer Model

Since some of the skills may be potentially contained in a phrase consisting of several words, it is important to utilise a tool that is able to capture the semantic meaning of a phrase as a whole. In order to do so, for this research paper, based on the findings from our previous work (Riabchenko A., Zheleiko I., 2022) the Sentence Transformer model on the basis of SBERT was utilised. This is an open-source family of models that is able to be custom trained and fine-tuned for the needs of the research. However, as is the case with the majority of neural networks, the amount of data needed in order to train such a model is extensive. For a sentence transformer type of model that is especially difficult as due to the specification of the model training process one should feed into it already collected triplets of sentences with their similarity scores all ready available. Hence, due to this limitation furtherly described in the limitation section, for the purposes of this work pre-trained openly available on the sbert.net (Sbert, 2022) models were looked into.

In the work of Hertling S., Portisch J., and Paulheim H. (2022) the authors had compared several available pre-trained SBERT-based models: all-MiniLM-L6-v2, paraphrase-albert-small-v2, paraphrase-TinyBERT-L6-v2, paraphrase-mpnet-base-v2, paraphrase-MiniLM-L6-v2, paraphrase-MiniLM-L3-v2, all-mpnet-base-v2, all-distilroberta-v5. They found, that despite all models scoring exceptionally high in terms of recall, the all-MiniLM-L6-v2 was ultimately the best performing one. Such high metrics were sufficient to use the models as it is the need for pre-tuning those was not justified.

For the purposes of this paper the model "all-MiniLM-L6-v2" is applicable as well, as it is well rounded model that was trained on a significant corpus of texts raging from comments on such webpages as reddit to scientific papers, allowing it to be applied to a wide range of texts, despite their sources, grammar constructions and lexicon used, in total utilizing more than a billion of training tuples of sentences (Hugging face, 2022).

In order to optimize the code, first of all, the embeddings for the skills described in the utilized taxonomy were obtained and kept in the separate dataset. After that a cycle was created that would go through all the phrases in the educational courses' descriptions, encode them and retain cosine-similarity scores through the in-build function "util.cos_sim". From all the found similarity scores only the pairs containing value larger than a specified threshold (in more detail the threshold value is described in the Model's Performance

Assessment section of the paper) were kept, with others being marked as "new" to the taxonomy as they do not have a close enough semantically match in the taxonomy to be labeled with an existing competency. Following this, based on the manual assessment of the extracted phrases, up to 4 matches with the highest similarity scores were chosen, offering the highest probable semantic similarity between a skill mentioned in the educational program text and a competency found in the utilized taxonomy.

3.4 Proposed Model's Performance Assessment

Based on the methodology utilized in the works of (Sibarani E. M., Scerri S., etc., 2017) and (Ayadi. A., Auffan M. and Rose J., 2020) so as to evaluate the model's goodness of fit, a manually curated "golden corpus" of labelled phrases was created. In order to develop such a corpus, 20 percent of educational courses texts were randomly pooled from the main text file, with each phrase being divided into key-phrases through the application of RAKE and manually labeled with the related competences found within the utilized taxonomy. If a certain concept within the test data could not be associated with any of the skills from the taxonomy, it was labeled as "new", and if less than four competences had been extracted within the phrase, the rest of the cells are empty Meaning, that "new" denotes an extracted competency which cosine similarity score with any of the competences existing in the taxonymy is less than the set threshold value (several threshold values were tested and associated F-score values calculated. More information on that can be found later in the text), while empty cells mean that in the phrase less than four competencies was found. Four was an arbitrary value, as it denotes the maximum amount of competences that was found in all of the extracted phrases. However, in most of the labelled cases less than four competencies were identified while proceeding with manual labelling of the Golden Corpus.

Phrase	Competence 1	Competence 2	Competence 3	Competence 4
assess their flexibility and openness to new ideas to inspire other team members and to create and sustain a positive, productive atmosphere.	Teamwork	Openness to the new	Influence on others	Training and development of others

communicating effectively in written and verbal forms.	Verbal Communications	Written Communications (Letter)		
optimize results while managing the constraints and stakeholder communications.	Result orientation	Assessing risks, constrains and business readiness to change	Planning and organization of communications with stakeholders	

Table 2. Extract from the golden corpus.

The performance of the ‘all-MiniLM-L6-v2’ model was evaluated with the following metrics: recall, precision and F-score (Himmelhuber A., Grimm S., Runkler T., and Zillner S., 2021), Gugnani, Akshay & Misra, Hemant. (2020) (Ptiček M., 2021). In order to enable calculation of these metrics we first need to obtain confusion matrix elements, such as:

TP (true positive)	The model provided to phrase the same label as the one found in the golden corpus
TN (true negative)	The model labelled a phrase as a “new” when in the golden corpus it is labelled as “new” as well. If the model did not retrieve any label and no label was given in the golden corpus it constitutes as TN as well.
FP (false positive)	The label retrieved by the model does not match the label given in the golden corpus
FN (false negative)	The model labelled a phrase as “new” when in golden corpus there is a different label

Table 3. Detailed description of the confusion matrix’s elements.

The evaluation of the confusion matrix elements is done in the following fashion: for each of the extracted phrase in the golden corpus up to 4 labels are manually assigned (if the skill is relevant to the taxonomy it is marked as a competence from it, if not it is marked as new), and are collected in the set GC_i, which can look as (“Competence from taxonomy 1”, “Competence from taxonomy 2”, Competence from taxonomy 3”, “new”). Then another set is created, M_i, comprised of the labels identified for the same phrases, in this case though, by the model; it may look like (“Competence from model 1”, “Competence from model 2”, “new”, “new”). The following step is the calculation of confusion matrix elements for each phrase i in the golden corpus: $TP_i+TN_i+FP_i+FN_i=4$.

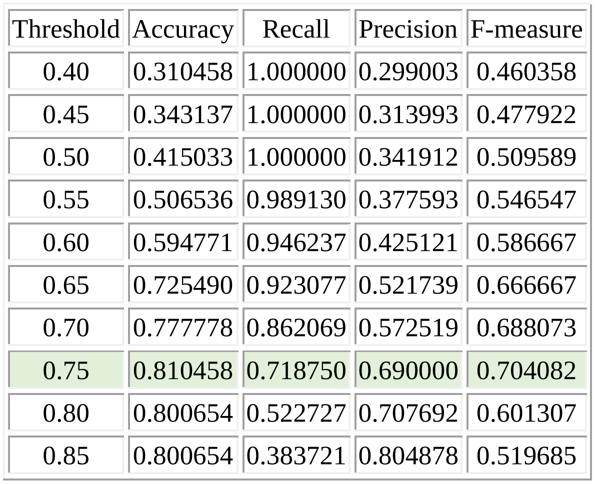
Accuracy allows us to see, how efficient our classification is. It is calculated as the ratio between all correctly identified elements divided by the sum of all classified elements (TP+TN+FP+FN).

Recall shows, how well the model is able to rightly classify an object. It is estimated as a ratio between TP and the sum of TP and FN (objects classified correctly divided by the number of all objects classified as relative).

With precision we can see, how many labelled objects were labelled correctly: ratio between TP and the sum of TP and FP (correctly labelled objects to the total number of all labelled objects).

As one may notice, there is a presence of a certain trade-off between precision and recall. In order to account for that F-score can be used, as it is a harmonic mean between these two values. F-measure value ranges between 0 and 1, where 0 identifies a situation whether either recall or precision are equal to 0 and 1 when both of the values achieve 100%. Due to the main interest of the current paper being the model's ability to correctly classify skills, the F-score would be used as the most important metric.

In order to find the most efficient model's configuration, a range of thresholds was tested from 0.4 to 0.9 with a step of 0.05. The threshold was determining whether the cosine similarity score between a phrase from the educational program text and a competency from the taxonomy is high enough to use the latter as a label for the extracted text or should it rather be labeled as "new", previously unseen concept in the taxonomy. As can be seen in the table 4, the highest F-measure score was achieved with the threshold value equal to 0,75 yielding the F-score of 70% which is a relatively high result considering the fact that it was not a binary classification task, and the number of possible labels is near 300. In their study, Himmelhuber A., Grimm S., Runkler T., and Zillner S., 2021 argue that their achieved metrics of recall being equal to 1 and precision being as low as 0.15 (which would yield only 26.1% f-score) is already enough to justify the automatization of the process as compared to manual labelling from scratch and will save both on labour time and expenses associated with the need of expertise work. Hence this was the value chosen for the final model specification.



Threshold	Accuracy	Recall	Precision	F-measure
0.40	0.310458	1.000000	0.299003	0.460358
0.45	0.343137	1.000000	0.313993	0.477922
0.50	0.415033	1.000000	0.341912	0.509589
0.55	0.506536	0.989130	0.377593	0.546547
0.60	0.594771	0.946237	0.425121	0.586667
0.65	0.725490	0.923077	0.521739	0.666667
0.70	0.777778	0.862069	0.572519	0.688073
0.75	0.810458	0.718750	0.690000	0.704082
0.80	0.800654	0.522727	0.707692	0.601307
0.85	0.800654	0.383721	0.804878	0.519685

Table 4. Model’s performance metrics for different threshold values.

Within the table 5 you can see the exempt of comparison between the skills labels resulting from the algorithm application to the randomly extracted phrases and labels assigned manually with assigned elements of confusion matrix.

Phrase from course description text	Competencies identified by model				Manually assigned labels			
	Comp 1	Comp 2	Comp 3	Comp 4	Comp 1	Comp 2	Comp 3	Comp 4
you will get a solid introduction to for example machine learning and neural networks, and you will learn where and how ai methods are applied in real life.	machine learning (TP)	artificial intelligence (TP)	new (TN)	(TN)	machine learning	artificial intelligence	new	(TN)
students can put the presented models and analysis methods into use with matlab or excel, where applicable, and solve real-life decision-making problems using the methods.	ms excel (TP)	decision-making (TP)	new (FN)	(TN)	ms excel	decision-making	methods and standards of business analysis	(TN)

practical programming in python.	Python (TP)	(TN)	(TN)	(TN)	python	(TN)	(TN)	(TN)

Table 5. Extract from the table comparing output by the ST model and labels within Golden Corpus.

Following that, the specification of the algorithm was applied to the rest of the data. The identified competences for the phrases were aggregated to sentences, and then aggregated again to be representative of the educational course's text description. Any of the duplicates were deleted.

3.5 Assessment of the open university courses analyzed

From the obtained data with all educational courses being labeled one can get following insights. In the educational market about 30% of the skills described in the taxonomy were found. Amongst these found skills the most prominent are hard skills. These were found in almost 70% of educational courses offered by open universities. These are closely followed by soft skills, constituting about 19% of the total set.

Hard skills were mainly focused around developing attendees' strategic capabilities (strategic management, strategic planning and strategic thinking competencies were mentioned in around one third of all educational courses), business administration-related competencies such as innovation management, knowledge in the area of marketing & sales, risk management and skills oriented at stakeholder management and engagement. In terms of the hard skills associated with more technical, evaluative skills the most popular was Data analysis, offered in 22 courses with the closest follow-up being Machine learning, found in 6 educational courses.

Finnish open universities also pay attention to developing learners' soft skills as well, with the most widespread being ones related to working with others: verbal communications (found in 18 educational courses), ethics (17 courses), leadership (15 courses), building relationships (14 courses) and ability to have influence on others (9 courses). When it comes to more individual's characteristics, here the most popular are decision-making (13 courses), responsibility (9 courses) and creativity (creative problem solving and creative thinking) – found in 8 educational courses.

Products, knowledge areas, tools, constitute almost 10% of the set. The most popular products proficiency in which is thought are Windows office products: MS Excel, offered in 5 educational courses, MS PowerPoint – 2 courses, MS Project, Archi (the only product that is not a part of the Windows 365) and MS Word were offered in one course each. When it comes to tools, the most popular were Python (found in four educational courses), Java (found in two educational courses) and SQL (offered in two courses as well). Only one framework mentioned in the taxonomy was taught in the form of an open university: Business model canvas.

When compared to the core competencies discovered in our previous research (Riabchenko A., Zheleiko I., 2022) for the position of the business analyst, one may see that not all of the most in-demand skills (extracted from the job description data collected from the currently found on the recruiting platforms such as “Monster.com” and end work-related online service “LinkedIn”) are currently (2022 autumn semester, 2023 spring semester) being covered by the offered by open universities educational courses. The usage of such product as Business Studio is not being taught in any of the current educational courses and could be potentially added to the curriculum to make sure that attendees will acquire skills that would be in-demand within current job market.

Competency	Number of occurrences in JDs	Number of courses teaching it
Requirements identification and management	70	10
Teamwork	64	3
Verbal communications	60	18
Development Project Management	60	6
Project Management Systems	59	7
Customer experience design	56	2
Learning ability	54	1
Business Process Design	51	1
Business Studio	50	0
Data Analysis	48	22
Business Intelligence Management	47	2

System Testing	46	4
Ms Project	44	1
Development of technical specification for the system	44	1
Initial Project planning	43	4
Analysis and processing of business capabilities	43	0
BI (Business intelligence)	40	2
Project Budget Management	40	3
Modeling and analysis of existing business processes of the organization	38	4
Formation and management of the project team	38	0
Requirements management (prioritization, tracking etc.)	37	3
Gathering information about business problems or business opportunities	34	1
Process mining	33	1
Development of a technical project system	33	1
MS SQL	33	2
Development of system test scenarios	31	0
Management based on stakeholders and their concerns	31	7
IT Project Management	31	3

Table 6. Comparison of core competencies for Business Analysts as derived from currently in demand job description data.

3.6 Matching knowledge gap and educational course

In order to find suitable courses, completion of which would fill the knowledge gap, first of all, only those educational courses are filtered out that contain the skill in question. After that they are ranked according to their match rank value, which is calculated as the ratio between the number of matches in the search term and skills contained by the educational program and total amount of potential matches (estimated as the longest length between the search term and the competencies contained within an educational program).

$$Match\ Rank = \frac{Number\ of\ matches\ between\ the\ search\ term\ and\ couse's\ skills}{Total\ number\ of\ potential\ matches}$$

Formula 3. Calculation of Match Rank

As an input either one skills is provided, or a set of skills. For example, for the search term of “Machine learning” the following courses would be showed. It is worth mentioning that “new” competences, that were not closely related to the utilized taxonomy, were included based on the assumption, that the more skills the course covers, the less detailed they are. Therefore, for example “Pattern Recognition and Machine Learning” course’s Match Rank is not 1, but rather 0.5, since big part of the course is Pattern Recognition which was not mentioned in the taxonomy, but plays an important role in the course’s contents.

Course name	Identified skills	Calculated Match Rank
Pattern Recognition and Machine Learning	'new', ' Machine Learning'	1/2=0.5
Advanced Data Analysis and Machine Learning	'new', ' Machine Learning', ' Data analysis'	1/3=0.33
Artificial Intelligence and Circular Economy	'new', ' Machine Learning', ' Artificial intelligence', ' AI'	1/4=0.25
Data Analysis with Python 2021	'new', ' Python', ' Machine Learning', ' Data analysis',	1/4=0.25
Building AI	'new', ' Python', ' Machine Learning', ' Artificial intelligence', ' AI'	1/5=0.2

Table 7. Match Rank calculation for one skill search term: {Machine learning}

In case of search consisting of several skills for example “Machine Learning” and “Python”, the ranking would look as follows in the table below. You can see that courses that do contain in them are ranking higher. In the those, where python was not mentioned the priority is given to those with lesser number of competencies taught.

Course name	Identified skills	Calculated Match Rank
Data Analysis with Python 2021	'new', ' Python', ' Machine Learning', ' Data analysis',	2/4=0.5

Building AI	'new', ' Python', ' Machine Learning', ' Artificial intelligence', ' AI'	2/5=0.4
Pattern Recognition and Machine Learning	'new', ' Machine Learning'	1/3=0.33
Advanced Data Analysis and Machine Learning	'new', ' Machine Learning', ' Data analysis'	1/4=0.25
Artificial Intelligence and Circular Economy	'new', ' Machine Learning', ' Artificial intelligence', ' AI'	1/5=0.20

Table 8. Match Rank calculation for several skills search term: {Machine Learning, Python}

After the match rank is calculated, the final output of the program looks as follows. The user will see sorted courses, showing in the decreasing order of the match rank. The provided information in the output is the city, where university is based, name of the educational platform/university, title of the course, calculated match rank and an excerpt of the course’s description.

City	Platform	Course Title	Match Rank	Program Description text exempt
Lappeenranta	LUT open university courses	Pattern Recognition and Machine Learning	0.5	select appropriate pattern recognition and machine learning methods and implement a working solution for a specific problem. introduction to pattern recognition, supervised, semi-supervised and unsupervised machine learning. feature processing and selection, system evaluation. statistical pattern recognition, bayesian inference and parameter estimation. Linear and non-linear classifiers based on artificial neural networks and support vector machines. Context-dependent and reinforcement learning. semi-supervised, unsupervised and method-independent learning
Lappeenranta	LUT open university courses	Advanced Data Analysis and Machine Learning	0.33	the student will be able to pre-process, visualise and analyse multivariate synthetic and real-world data. understand and use state-of-the-art regression methods, probabilistic graphical models and deep neural networks. data characteristics and pre-processing, linear and nonlinear dimensionality reduction. Logistic, principal component and advanced regres-

				sion methods. probabilistic graphical models, bayesian networks and probabilistic inference. deep and convolutional neural networks. case-based topics on data analysis and machine learning
Kuopio	University of Eastern Finland	Artificial Intelligence and Circular Economy	0.25	describe what are artificial intelligence (ai) and machine learning (ml), classify different ml algorithms, explain the basic principles of each algorithm, express the advantages of ai and ml. apply the ml algorithms in practice with software. predict the results with given data package. demonstrate the given successful cases of ai applications in circular economy. interpret and illustrate how ai can be used in sustainable business to advance the transition of a circular economy. artificial intelligence.
Helsinki	Mooc.fi	Data Analysis with Python 2021	0.25	in this course an overview is given of different phases of the data analysis pipeline using python and its data analysis ecosystem. useful tool for data analysis is machine learning, where a mathematical or statistical model is fitted to the data. these models can then be used to make predictions of new data, or can be used to explain or describe the current data.
Helsinki	Mooc.fi	Building AI	0.20	you will get a solid introduction to machine learning and neural networks, and you will learn where and how ai methods are applied in real life. as a result of this course, you will be able to craft your own ai idea and present it to the community.

Table 9. Output of recommended courses for the search term “Machine Learning”.

Validation of the matching task is more of a difficult task. The most thematically closest to the current thesis work is Shakya A. and Paudel S. (2019), where authors were developing a job description-job seeker matching algorithm through the ESCO ontology. In their research paper, in order to validate the proposed ranking algorithm, Shakya A. and Paudel S.had asked 7 human evaluators, each with at least 5 years’ worth of expertise in IT to validate the output of the proposed algorithm. In the case of the current research question of matching educational courses to skills, perhaps a HR recruiter with expertise in the IT could be asked to validate the provided by the algorithm recommendations.

3.7 Summary of Chapter 3

In the chapter 3 the concepts described in the chapter 2 were applied to the collected data. To begin with, the general justification of the proposed method was outlined. Next, the proposed model was applied to 20 percent of randomly extracted phrases of the main data frame and manually labeled with relevant skills found within the taxonomy. In order to evaluate the model's goodness of fit such metrics as recall, precision and F-score were used. After the comparison of different model specifications, the best performing one was chosen and applied to the main data frame. The achieved F-score value was high enough to deem the model successful based on the proposed requirements. Therefore, meaning that the proposed model can indeed be applied to the educational program text descriptions in order to extract skills. What is more, a brief analysis of the educational market saturation with IT and managerial related competencies was provided. Moreover, methods of accessing both semantic labelling and matching of the courses and skills were outlined.

4 Managerial applications of the constructed algorithm

As was stated before, the main purpose of the current research was offering a potential outline of an algorithm that would be able to analyze current situation on the open universities' educational courses market and could be used as a tool, swiftly providing a list of ranked most suitable courses based on the skill of interest.

In more detail, one can find the proposed model and analysis useful in the current areas:

- The companies can use this tool in order to optimize the training process of employees working within a IT-related team, by quicky accessing the courses available in the finish educational market. When fed a certain knowledge gap or a list of knowledge gaps, the algorithm retrieves ranked courses suggestions, with names of the courses, universities and cities where they are based.

- One may use the algorithm to very quickly assess the course contents and almost immediately see the key competences as identified by utilized aggregated taxonomy that they will obtain after completion of the course.
- The proposed analysis offers insights on the current state of the open university educational courses, suggesting what competencies are covered in several of universities and courses and where the supply maybe overly saturated by competitors and what courses are not as covered. This information may be of use for universities and companies developing educational courses to find a niche/competency that is currently not occupied and develop relevant course, covering it. What is more, since the algorithm is applicable to job description data, universities may use information on the current skills in demand derived from the job advertisements and compare them with which are currently covered on the educational market and what is not.
- Both in the current thesis and the previous research the all-MiniLM-L6-v2 model was proven to be applicable to process texts related to the IT-related competences including educational program texts, therefore it can be applied to a larger variety of texts: such as universities degree program descriptions, educational platforms courses descriptions that are offered by businesses, job advertisement data and CVs of protentional candidates.

5 Limitation of the study and ideas for future research

When it comes to machine learning, especially the text related subfields of it, training of the models requires immense amount of information, for example the pre-trained model utilized in this thesis (‘all-MiniLM-L6-v2’) was trained on nearly a billion of training tuples of sentences (Hugging face, 2022). The collection of so much date requires time and resources and training the model on them requires time, resources and computational power, the usage of which were limited in the case of the current thesis. However, it might be promising to train a language SBERT-like model on the competences related data to make it more adjusted to the specifics of labour market related lexicon.

In the current thesis a very basic view on the ranking of educational courses through the implementation of semantic matching was outlined. The main concern for the matching was semantic relevance of the identified skills with the skill being searched. However, with acquiring more data, a more detailed matching algorithm could be used. Including pricing, length of the courses etc.

Courses of some open universities were present in StudyInfo (StudyInfo. 2022) and Opintopolku.fi but some of the universities were not present there and their educational courses had to be retrieved from their respective web sites. What is more, these own web sites were not organised in the same fashion, Moreover, the choice of the subject fields was varying in universities with some universities offering technical courses for example, Tampere university (Tuni, 2022), while some offered mostly soft skill related courses, like Haaga-Helia (Haaga-Helia, 2022). As a result, the development of a unifying web crawler was a task out of the scope of the current thesis. A creation of an aggregator (alike existing services concerning job description data, such as Skills OVATE, Skills intelligence etc.) of all open courses text descriptions seems to be a next logical step in this research, which would facilitate the analysis of the skills being covered by the existing educational platforms.

The current thesis was an example of utilization NLP techniques for semantic labelling of educational courses text descriptions based on the case of Finland’s offer of open universities’ courses concerning IT, management and business- related competencies. However the method could have been potentially expanded to other than IT-related industries, given that the taxonomy for the characteristic competency is utilized.

6 Conclusion

The main purpose of the current paper was development of a recommendation providing algorithm that would upon feeding into it either a skill or a set of skills, retrieve educational curses, completion of which could potentially cover the existing knowledge gap. In order to do so, the data on the contents and competencies to be learned was collected from StudyInfo and Opintopolku.fi as well as universities’ own web pages.

The proposed algorithm was first tried on the specifically developed golden corpus with randomly extracted phrases from the educational courses texts manually labelled with taxonomy's competencies and several model's specifications were run in order to achieve the highest possible model's performance, which quality was achieved through the usage of such evaluation metrics as accuracy, precision, recall and F-score. The highest F-score (70,4%) was achieved with the threshold value of 0.75. Later this value had been applied to the algorithm running on the whole data frame. Ensuring that indeed the NLP methods, such as Sentence Transformer language model can be applied to extract competencies from educational course text descriptions.

After the subjecting the data to the Sentence Transformer model, labels for phrases were obtained. Later these labels were used as a filter value for educational courses. When the user inputs a certain knowledge gap (a skill needed to be learned) the algorithm retrieves a list of ranked relevant educational courses, completion of which will help with filling the knowledge gap.

What is more, a brief analysis of the current educational courses market was offered, giving insights on the market's saturation for certain skills. Moreover, a comparison of those with core competencies that are currently in demand for the position of a data analyst was done, showing which competencies are currently not to well covered by the existing curriculum, answering the second research question.

References

Alake R., 2020. *Understanding Cosine Similarity And Its Application.* Available at https://towardsdatascience.com/understanding-cosine-similarity-and-its-application-fd42f585296a. (Accessed: 09 Oct. 2022).

Anthony S., 2022. *What Is a Skills Taxonomy and Why Do You Need It?* Available at https://www.linkedin.com/business/talent/blog/learning-and-development/what-is-a-skills-taxonomy-and-why-do-you-need-it (Accessed 02 September 2022)

Bayraktar O., Ataç C. 2018. The Effects of Industry 4.0 on Human Resources Management. Globalization, Institutions and Socio-Economic Performance (pp.337-360)

Berragan C., Singleton A., Calafiore A. & Morley J. (2022). Transformer based named entity recognition for place name extraction from unstructured text. International Journal of Geographical Information Science, pp. 1-20.

Bian S., Zhao W. X., Song Y., Zhang T. and Wen J. Domain Adaptation for Person-Job Fit with Transferable Deep Global Match Network. 2019. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Pp. 4810–4820.

Bonaros B., 2022. Mastering Sentence Transformers For Sentence Similarity. Available at: https://predictivehacks.com/mastering-sentence-transformers-for-sentence-similarity/. (Accessed: 31 Oct 2022).

Bondielli and F. Marcelloni, “On the use of summarization and transformer architectures for profiling résumés,” Expert Systems with Applications, vol. 184, 2021

Brownlee J., 2020. 4 Distance Measures for Machine Learning. Available at: https://machinelearningmastery.com/distance-measures-for-machine-learning/ (Accessed: 18 Novemer 2022).

Cedefop (2022). Skills-OVATE. Available at: https://www.cedefop.europa.eu /en/tools/skills-online-vacancies (Accessed: 16 September 2022)

Cedefop (2022). Cedefop and Eurostat formalise joint approach to online job advertisement data. Available at: https://www.cedefop.europa.eu/en/news/cedefop-and-eurostat-formalise-joint-approach-online-job-advertisement-data (Accessed: 16 September 2022)

Cedefop (2021). Understanding technological change and skill needs: big data and artificial intelligence methods. Cedefop practical guide 2. Luxembourg: Publications Office. Available at: http://data.europa.eu/doi/10.2801/144881 (Accessed: 16 September 2022)

Chan C.S., Pethe C. and Skiena S. Natural language processing versus rule-based text analysis: Comparing BERT score and readability indices to predict crowdfunding outcomes. 2021. Journal of Business Venturing Insights, 16.

Coats D., 2022. Semantic Search: How It Works & Who It's For. Available at: https://www.searchenginejournal.com/semantic-search-how-it-works-who-its-for/438960/#close (Accessed: 23 October 2022)

Coursera (2022). Available at: https://www.coursera.org/ (Accessed: 07 September 2022)

Custom Insight (2022). What is 360 Degree Feedback? Available at: https://www.custominsight.com/360-degree-feedback/what-is-360-degree-feedback.asp (Accessed: 07 September 2022)

Dadzie A., Sibarani E., Novalija I., and S. Scerri. Structuring visual exploratory analysis of skill demand. Journal of Web Semantics, dec 2017.

Devlin J., Chang M., Lee K., and Toutanova K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Digital City Planner (2022). Available at: https://digicityplanner.com/

Djumalieva J., Sleeman C., 2018. An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts. ESCoE Discussion Paper 2018-13.

Ministry for Foreign Affairs, Ministry of Education and Culture 2017. *Education In Finland.* Key to the nation's success, brochure, Otavamedia OMA, Helsinki

Efremova N., Shapovalova O., Huseynova A., 2020. Innovative technologies for the formation and assessment of competencies and skills in the XXI century. published by EDP Sciences. E3S Web of Conferences 210, 18021 (2020)

Fazel-Zarandi, M., & Fox, M.S. (2009). Semantic Matchmaking for Job Recruitment: An Ontology-Based Hybrid Approach.

FEAPO (2018). The Guide to Careers in enterprise Architecture. https://feapo.org/wp-content/uploads/2018/10/Guide-to-Careers-in-Enterprise-Architecture-v0.5-clean-copy-copy.pdf (Accessed: 05 September 2022)

Gate (2022). GATE: General Architecture for text engineering. Available at: https://gate.ac.uk/ (Accessed: 09 September 2022)

Gugnani, Akshay & Misra, Hemant. (2020). Implicit Skills Extraction Using Document Embedding and Its Use in Job Recommendation

Gruetzemacher R., Paradice D., Deep Transfer Learning & Beyond: Transformer Language Models in Information Systems Research. 2022. ACM journal, Vol. 54., No. 105. Pp 1-35.

Grüger, Joscha & Dr. Schneider, Georg. (2019). Automated Analysis of Job Requirements for Computer Scientists in Online Job Advertisements. 226-233

Haaga-Helia (2022). Avoin AMK: Koulutustarjonta. Available at: https://www.haaga-helia.fi/en/open-uas (Accessed: 09 September 2022).

Han J., Kamber M. and Pei J. (2012). 'Getting to Know Your data', in The Morgan Kaufmann Series in Data Management Systems. *Data Mining*.

Hertling S., Portisch J., and Paulheim H. (2022). KERMIT – A Transformer-Based Approach for Knowledge Graph Matching. Morgan Kaufmann. Pages 39-82.

Himmelhuber A., Grimm S., Runkler T., and Zillner S. 2021. Ontology-Based Skill Description Learning for Flexible Production Systems

Holt J. and Perry S. (2011) A pragmatic guide to competency tools, frameworks and assessment, BCS The Chartered Institute for IT London.

Hugging face (2022). All-MimiLM-L6-v2 model. Available at: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 (Accessed: 07 November, 2022)

IBM (2022). What is industry 4.0? Available at: https://www.ibm.com/topics/industry-4-0 (Accessed: 05 November 2022)

International Institute of Business Analysis (2017). *Business Analysis Competency Model*, International Institute of Business Analysis, Toronto, Canada.

Kamran M., Anjum M. S. 2017. The use of Ontologies for Eective Knowledge Modelling and Information Retrieval. Applied Computing and Informatics.

Kandi M., Language Modelling for Handling Out-of-Vocabulary Words in Natural Language Processing. PhD thesis, 08 2018

Manatal (2022) 'Revolutionizing HR: Natural Language Processing in Recruitment'. Available at: https://www.manatal.com/blog/natural-language-processing-in-recruitment (Accessed 01 October 2022)

Merchant Q. (2021) 'How to evaluate competency in employees', Business Finance Articles. Available at: https://businessfinancearticles.org/how-to-evaluate-competency-in-employees (Accessed 07 September 2022)

Mestrovi A., Cal A., An ontology-based approach to information retrieval, in: In Semantic Keyword-based Search on Structured Data Sources, Springer, 2016, pp. (pp. 150–156)

Mhamdi D., Moulouki R., M. E. Ghoumari, M. Azzouazi, and L. Moussaid, "Job recommendation based on job profile clustering and job seeker behavior," Procedia Computer Science, vol. 175, pp. 695–699, 2020

Mokhtari, N.I. (2022). What are Siamese Neural Networks in Deep Learning? Towards Data Science Available at: https://towardsdatascience.com/what-are-siamese-neural-networks-in-deep-learning-bb092f749dcb (Accessed at 05 November 2022)

Mooc.fi (2022). Available at: https://www.mooc.fi/en/#courses (Accessed: 09 September 2022)

Open AI (2022). ChatGPT: Optimizing Language Models for Dialogue. Available at: https://openai.com/blog/chatgpt/ (Accessed 05 December 2022).

Open Group (2006). TOGAF® Series Guide Architecture Skills Framework. Available at: https://pubs.opengroup.org/architecture/togaf8-doc/arch/chap30.html (Accessed: 04 September 2022)

Opintopolku (2022). Avoin yliopisto. Available at: https://opintopolku.fi/wp/yliopisto/avoin-yliopisto/ (Accessed: 20 September 2022)

Opintopolku (2022). Explore lifelong learning in Finland. Available at: https://opintopolku.fi/konfo/en (Accessed: 25th November 2022)

Otendo (2022). What is a Taxonomy? Available at: https://tendocom.com/glossary/taxonomy/ (Accessed 30 August 2022)

Otter D., Medina J. and Kalita K. 2019. A Survey of the Usages of Deep Learning for Natural Language Processing. Transactions On Neural Networks and Learning Systems, Vol. Xx, No. X.

Ptiček M. 2021. How good BERT based models are in sentiment analysis of Croatian tweets: comparison of four multilingual BERTs. Faculty of Organization and Informatics. University of Zagreb. 32nd CECIIS, October 13-15, 2021

Textrank: Bringing order into texts. In Proceedings of EMNLP 2004 (ed. Lin D and Wu D), pp. 404–411. Association for Computational Linguistics, Barcelona, Spain.

Tuni (2022). Tampere University. Available at: https://www.tuni.fi/en/search ?organisation=1&search_type=study&education_type=openuni&search_lang_instruction= en (Accessed: 09 September 2022)

Panoutsopoulos H., Brewster C. Data-driven Update of AGROVOC Using Agricultural Text Corpora

Rajashekharaiah K. M., Ganiger S. (2018). Comparative Study on Keyword Extraction Algorithms for Single Extractive Document.

Ramli F., Noah S., Kurniawan T., Ontology-based information retrieval for historical documents, in: Third International Conference on Information Retrieval and Knowledge Management (CAMP), IEEE, 2016, pp. (pp. 55–59)

Reimers N., Gurevych I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Ubiquitous Knowledge Processing Lab (UKP-TUDA). 11.

Reimers, N. (2022) ‘Pretrained Models - Sentence-Transformers documentation’, Sbert. Available at: https://www.sbert.net/docs/pretrained_models.html (Accessed: 07 November, 2022)

Rentzsch R., Staneva M. (2020). Skills-Matching and Skills Intelligence through curated and data-driven ontologies. Proceedings of the DELFI Workshops 2020, Heidelberg, Germany, September 14, 2020

Rose S., Engel D., Nick Cramer and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. Text Mining: Applications and Theory, John Wiley & Sons, Ltd

Sales Enablement (2019) Sales enablement analytics report 2019. Available at: https://salesenablement.pro/expertise/ (Accessed: 08 September 2022)

Seema S. The handbook of competency mapping: understanding, designing and implementing competency models in organizations. 2016.

Sibarani, Elisa & Scerri, Simon & Morales, Camilo & Auer, Sören & Collarana, Diego. (2017). Ontology-guided Job Market Demand Analysis: A Cross-Sectional Study for the Data Science field. Proceedings of the Semantics conference, Amsterdam, the Netherlands

Shakya A., Paudel S. Job-Candidate Matching using ESCO Ontology. (2019). Journal of the Institute of Engineering January 2019, Vol. 15 (No. 1): 1-13

Skillshare (2022). Available at: https://www.skillshare.com/ (Accessed: 07 September 2022)

Spacy (2022). Available trained pipelines for English. Available at: https://spacy.io/models/en (Accessed: 06 November 2022)

Stember J., Shalu H. Deep reinforcement learning with automated label extraction from clinical reports accurately classifies 3D MRI brain volumes. 2021.

Study Info (2022). Higher Education. Open Studies. Available at: https://studyinfo.fi/wp2/en/higher-education/open-studies/ (Accessed: 05th October 2022)

Sun, C.; Qiu, X.; Xu, Y.; Huang, X. (2019). How to Fine-Tune BERT for Text Classification?, Chinese Computational Linguistics-18th China National Conference, Kunming, China, 18–20

Symonds, C. (2022) ‘Tracking Employee Performance Metrics: All You Need to Know’, Factorial Blog. Available at: https://factorialhr.com/blog/employee-performance-metrics/ (Accessed: 07 September 2022)

Tenney T., Xia P., Chen B., Wang A., Poliak A., McCoy T., Kim N., Van Durme B., R. Bowman, Das D., and Pavlick E. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. Published as a conference paper at ICLR 2019.

Tortorella, G., Miorando, R., Caiado , R., Nascimento, D., Staudacher, A.P., The mediating effect of employees’ involvement on the relationship between Industry 4.0 and operational performance improvement, Total Quality Management & Business Excellence vol 29, 2018

Vukajlović D., Brzaković M., Assessment of employees competences carried out by different management levels. 2016. Ekonomika No. 62(3). Pp: 47-56.

World Economic Forum. Towards a Reskilling Revolution: A Future of Jobs for All. January 2018.

Wunderlich J., Tilebein M. Transforming Intended Learning Outcomes expressed in natural language into elements of an ontology. Towards the formalisation of Intended Learning Outcomes for use in Curriculum Maps. Proceedings of DELFI Workshops 2019, Berlin, Germany

Yasar, M.F., Ünal, Ö.F., Zaim, H., Analyzing the Effects of Individual Competences on Performance: A field study in services Industries in Turkey, Journal of Global Strategic Management 2(7), 2013, p. 67-81

Zippia (2022). Available at: https://www.zippia.com/ (Accessed: 14 September 2022)

Zhang Y., Yang C., and Niu Z. A research of job recommendation system based on collaborative filtering. 2014. In Proceedings of the 7th International Symposium on Computational Intelligence and Design. Pp. 533–538.

Zhu C., Zhu H., Xiong H., Ma C., Xie F., Ding P., and Pan Li. Person-job fit: Adapting the right talent for the right job with joint representation learning. 2018. ACM Transactions on Management Information Systems.