Automatic definition of engineer archetypes: A text mining approach

Computers in Industry 152 (2023) 103996

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

Automatic definition of engineer archetypes: A text mining approach

Francesco Lupi [a,*], Mohammed M. Mabkhot [b], Eleonora Boffa [c], Pedro Ferreira [b], Dario Antonelli [d], Antonio Maffei [c], Niels Lohse [b], Michele Lanzetta [e]

[a] Department of Information Engineering, University of Pisa, Italy
[b] Intelligent Automation Centre, The Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough LE11 3TU, UK
[c] Department of Production Engineering, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden
[d] Department of Management and Production Engineering, Politecnico di Torino, 10129 Torino, Italy
[e] Department of Civil and Industrial Engineering, University of Pisa, 56122 Pisa, Italy

ARTICLE INFO

Keywords:
Text mining
Engineering
Professional profile
Archetype
Latent dirichlet allocation
Industry 4.0

ABSTRACT

With the rapid and continuous advancements in technology, as well as the constantly evolving competences required in the field of engineering, there is a critical need for the harmonization and unification of engineering professional figures or archetypes. The current limitations in tymely defining and updating engineers' archetypes are attributed to the absence of a structured and automated approach for processing educational and occupational data sources that evolve over time. This study aims to enhance the definition of professional figures in engineering by automating archetype definitions through text mining and adopting a more objective and structured methodology based on topic modeling. This will expand the use of archetypes as a common language, bridging the gap between educational and occupational frameworks by providing a unified and up-to-date engineering professional figure tailored to a specific period, specialization type, and level. We validate the automatically defined industrial engineer archetype against our previously manually defined profile.

1. Introduction

In recent years, the significance of harmonizing, unifying, and regularly updating the comprehensive range of skills and competences that make up the professional figures has become increasingly evident (Fareri et al., 2020). In this context, the rapid technological progress and the mounting emphasis on sustainability issues necessitate that the professional figures of engineers remain up to date with the latest practices (Mabkhot et al., 2021). Consequently, engineer archetypes have recently been conceptualized as abstract and generic summarization identikits of a target engineer figure, compiled from a set of technical skills acquired through educational learning outcomes and occupational frameworks (Lupi et al., 2022).

The abstract representation provided by archetypes disclose an enormous potential in several activities such as *design* (i.e., design public or private education systems in accordance to a given archetype); *benchmarking* (i.e., identify the gap between the learning outcomes of proposed courses and the archetype); *assessment*, (i.e., understand the impact of a specific archetype on particular issues); *communication* (i.e., provide a comprehensive but summarized format of a specific archetype to stakeholders); and *harmonization* (i.e., push the boundary gaps between the educational and occupational frameworks and harmonize it with the industrial needs to benefit from the technological advancement).

Although the concept of archetypes has been recently introduced into the field of engineering from a novel and original perspective, defining the engineer archetypes remains a qualitative process that involves identifying and categorizing the technical skills possessed by professional figures through the use of expert panels. This is a time-consuming process that is susceptible to biases and challenges in comprehensively analyzing non-structured and multi-dimensional data from diverse sources (Lupi et al., 2022).

The utilization of text mining for defining engineer archetypes may present a more objective and efficient approach remarkably decreasing the time, expertise and resources required for manual methods. In detail, text mining can automatically extract the main engineering competences by processing large amounts of textual data (i.e., university courses and job descriptions), and finding trends and links between data, enhancing the precision of the output (Guo et al., 2016; Pejic-Bach et al., 2020). Considering these advantages, text mining provides a potent

* Corresponding author.
E-mail address: francesco.lupi@phd.unipi.it (F. Lupi).

https://doi.org/10.1016/j.compind.2023.103996
Received 18 April 2023; Received in revised form 15 July 2023; Accepted 28 July 2023
Available online 8 August 2023

technique for assisting experts in broadening the scope of the data to encompass more comprehensive and representative samples, leading to the identification of more unified archetypes. As a result, the experts' roles are elevated to assess and validate the results of text mining, ensuring the accuracy and relevance of the insights gained.

Taking into consideration the aforementioned, this research presents a technique capable of bringing professional engineering profiles into alignment and coherence via the utilization of automatic text mining methods for archetype definition. The input of the method is a raw corpus of textual descriptions of the professional figure under interest. The output is a catalog of keywords (i.e., technical contents and Bloom verbs), clustered in the main competences' groups, similarly to the representation proposed in (Lupi et al., 2022), using topic modelling algorithms.

The remainder of the paper is structured as follows: Section 2 presents the motivation, relevant literature, and related works on the concepts of archetype, text mining, and topic modelling. Section 3 describes the proposed methodology for the automatic definition of engineer archetypes. In Section 4, we validate the proposed method by comparing the automatically generated industrial archetype engineer with the benchmark manually defined in our previous paper. Finally, Sections 5 and 6 present the discussion and conclusion, respectively.

2. Background

The present section employs the funnel approach depicted in Fig. 1 to furnish the research background of this work. Specifically, it adopts a top-down approach, starting from the general and gradually narrowing down to the specific. Section 2.1 introduces the motivation behind the definition of archetypes. Section 2.2 serves as a comprehensive introduction to the essential concept of archetypes in the engineering field. This is followed by a discussion in the subsequent two sections on the text mining tools employed in the proposed methodology. Section 2.3 provides an outline of the topic modelling techniques while Section 2.4 provides an overview on the specific topic modelling algorithm adopted in this work (i.e., Latent Dirichlet Allocation (LDA)).

2.1. Archetype motivation

There exists a significant misalignment and gap between the educational and occupational realms. By aligning curriculum with industry needs and emphasizing practical experiences, we can better equip individuals to succeed in their chosen careers and ensure a smoother transition from education to the world of work.

As illustrated in Fig. 2 through the Venn diagram, within a specific professional domain, two sets of competences (c_i) exist. One set is associated with education side (E)= $\{c_i \mid i = 1, q,...t\}$ (i.e., supply), while the other set is related to occupation (O)= $\{c_i \mid i = 2, 3, h,...k\}$ (i.e., demand). In the current state of knowledge (Fig. 2a), the overlap between E and O can be observed, indicating that students who have completed their educational path are employed and meet market requirements (to varying degrees based on domain-specific factors such as field, geographic area, and time frame). Unfortunately, a clear, explicit, and automatic procedure for generating a list of competences belonging to the intersection $E \cap O$ is still absent, as indicated by the question mark (Fig. 2a). By automating the manual procedure for archetype definition presented in our previous work, two significant contributions are achieved, as depicted in the Fig. 2b, which represents an enriched knowledge state. The first (indirect) contribution involves enlarging the boundaries of the intersection between E and O through the application of the abstraction principle. This is accomplished by defining the archetype as a collection of n meta-classes (clusters or topics) to which competences belong. As a result, the perceived gap or mismatch between the occupational sides is reduced. For instance, a topic can encompass competences from both E and O (e.g., Topic1 T1 includes competences c_1 and c_2), or exclusively from O or E (Topic4 T4 includes c_2 and c_3). The second (direct) contribution involves the explicit definition of a structured list of competences in the form of keywords organized into those topics.

In conclusion, by considering the archetype as a set of topics (A) = $\{Tf \mid f = 1...6\}$, where Topicf (Tf) = $\{c_i \mid i = j...m\}$, it can be used to assess the current state of a curriculum and establish a feedback loop to add or remove specific competences from the program. Similarly, from a business managerial perspective, having knowledge of a shared and recognized archetype for a professional profile can provide insight and guide managers in making informed decisions during the selection and lifelong learning processes, bridging the gap with the current state of the art in the profession. As depicted in Fig. 2b, iterative adoption and updating of the archetype over time can be employed to evaluate the education and occupation domain. By comparing the current state to the desired archetype state, corrective actions can be taken to narrow the gap or reduce errors, thus pushing the limits of the E and O sets towards convergence and achieving a stable state where education and occupation are harmonized.

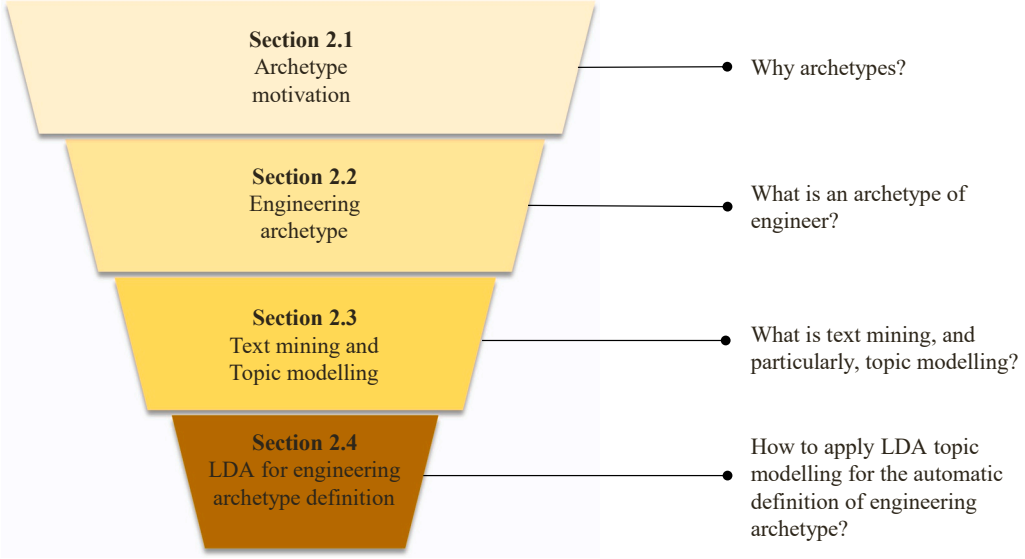

Fig. 1. Graphical representation of the funnel approach for the background of main concepts related to the current work.

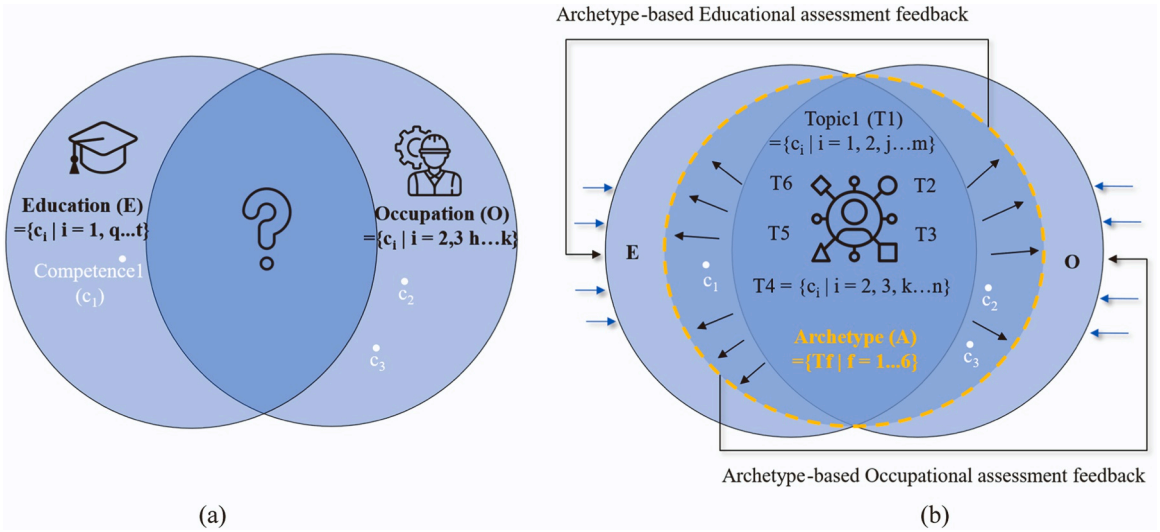

Fig. 2. Venn diagram of the competences ($c_i \mid i = 1…n$) within a specific domain. (a) Current knowledge state, where competences c_i can be categorized as educational-related {E}, occupational-related {O}, or both. (b) Desired knowledge state, where the actual overlapping of occupational and educational competences is explicit (i.e., list of competences, grouped by topics) and expanded (i.e., dotted yellow borders are pushed and pulled according to the arrows) by abstraction reasoning based on topic modelling (i.e., archetype's topics).

2.2. Engineering archetype

The archetype principle (i.e., abstraction) is inspired by the concept of personas, extensively utilized as an abstract representation of a singular, synthetic, and representative primitive model that mirrors the individuals comprising the target user (Karimova and Goby, 2020; Kong et al., 2018; Vincent and Blandford, 2014). The use of such abstraction has been advocated to summarize and facilitate communication of needs, improving the user experience by guiding teams in designing for generalized or meta users (Wood and Mattson, 2019; Floyd et al., 2008; Miaskiewicz and Kozar, 2011; Neate et al., 2019).

This concept has been also applied to the educational context to catch the motivation and learning strategy of students (Phuong et al., 2013). For instance, authors in (Phuong et al., 2013) states that a virtual student representing a student group similar in motivation and learning strategy to learn programming, enables the teacher to predict student behavior during the programming education course and design the course accordingly (Phuong et al., 2013).

Focusing on the engineering archetypes, both scientific research and regulatory documents sources (e.g., accreditation and standardization) provide different interpretation on what the engineering archetype is, and how it can be defined. Several sources refer to engineer professional profile instead of archetype. In the current work the two terms can be used interchangeably. Table 1 summarizes the retrieved literature, highlighting contribution and gap for each retrieved source.

According to Table 1, extant literature highlights the tight synergy between the educational and occupational worlds for defining professional engineering figures. The analyzed contributions provide a first attempt to establish a common definition of knowledge and competences for engineers by providing comprehensive frameworks for both education and training of engineers. Yet, such contributions provide generic and heterogeneous guidelines. Thus, further research is needed to harmonize the definition of engineering archetypes.

As for the methodologies, the analyzed literature lacks a structural approach for generating engineer archetypes. The input information is often based on manual surveys that may not capture all the knowledge, skills, and abilities essential for a particular engineering discipline. The existing regulatory documents attempt to map all the skills and abilities but lack to capture all variation across the different programs and institutions and are often limited to a specific geographical context. Therefore, this provides an incomplete picture of the required knowledge for a modern professional engineer. Finally, the analyzed contributions do not collect or process input data using automatic methods.

Given the above insights and gap, this study proposes a method to harmonize and unify professional engineering profiles or archetypes using automatic text-mining topic modelling techniques of heterogeneous (i.e., occupational and educational) data sources.

2.3. Text mining and topic modelling

Natural language processing (NLP) is a field that combines the power of computational linguistics, computer science, and artificial intelligence to enable machines to understand, analyze, and generate the meaning of unstructured data (e.g., natural human text and speech) (Chowdhary, 2020). The NLP approach usually involves the execution of a software pipeline composed of steps with the aim of extracting information from text, also known as text mining (Fareri et al., 2020).

Text mining for engineering involve the processing of textual data from a diverse range of engineering-related text sources, including policy documents (Massey et al., 2013), patents (Spreafico and Spreafico, 2021), social media (Chiarello et al., 2020), and research articles (Amado et al., 2018). The primary aim is to unveil hidden patterns, trends, and relationships within this information, providing businesses with valuable insights to inform decision-making, enhance products and services, and gain an edge in the market. Recently, text mining techniques have been heavily utilized to analyze Industry 4.0, identifying key themes historically discussed and tracking their development over time (Galati and Bigliardi, 2019), exploring the effects on job roles and skills (Fareri et al., 2020; Pejic-Bach et al., 2020) and compiling an enriched dictionary of relevant enabling technologies (Chiarello et al., 2018). Ultimately, the application of text mining had been demonstrated to support the engineering design process through the automatic extraction and interpretation of useful information from complex, open databases (Giordano et al., 2022).

Unstructured natural language text lacks a defined data model or meta level information, but often contains latent structure, including parts-of-speech, named entities, relationships between words, and word sense. This structure can be inferred through human interpretation or machine learning algorithms (Chen et al., 2016).

Among text mining approaches, topic modelling is an unsupervised task that captures hidden semantics in text. Techniques like Latent

Table 1
Engineering professional profile literature both from scientific publications and regulatory documents (e.g., accreditation and standardization) data sources.

Type and Ref	Contribution	Gap
Scientific article (Doré et al., 2021)	The work highlights a national study in Canada examining the development of undergraduate engineering students' identities. The study investigates students' understanding and prioritization of the attributes, variations across institutions and over time, and alignment with their definition of an engineer.	The paper only presents an overview of the study development and methods used, and no results or conclusions are presented yet.
Scientific article (Davis et al., 2005)	The paper outlines a process to define the profile of an engineer who performs well in professional practice. It includes technical, interpersonal, and professional skills, developed with input from both academic and non-academic engineers, to align with key roles in the field. The main professional profile features are identified (i.e., Comprehensive; Concise; Distinct; Organized; Action Orientation; Compelling).	The engineer profile is primarily based on self-reported data of 100 academic and non-academic engineers, which may be subject to bias.
Scientific article (Gainsburg et al., 2010)	The study constructed a "knowledge profile" from field observations of structural engineers to differentiate among engineering occupations and help university engineering programs better target and reflect the knowledge demands of the profession. The authors suggest that such profiles can help bridge the gap between formal education and practice-generated knowledge in engineering work.	The focus is more on mapping the different type of knowledge (education and practice generated) rather than capturing the identikit of structural engineering skills set and competences.
Scientific article (Lupi et al., 2022)	The authors proposed a general methodology to define specific engineer archetypes as a set of technical competences from educational Learning Outcomes and occupational European Skills, Competences, Qualifications, and Occupations (ESCO) frameworks. A collaborative expert-based working method and its application to the definition of industrial engineering is presented, extracting six main clusters of competences identifying the professional figure under interest.	The proposed methodology for defining archetypes can be improved by a more systematic and automated process for text data collection and analysis using text mining.
Standard regulation (CEAB)	The Canadian Engineering Accreditation Board (CEAB) Graduate Attributes (GAs) accreditation board proposes 12 attributes that describe the knowledge, skills, and attitudes that engineering graduates should possess upon completion of an accredited program in Canada. The CEAB GAs serve as a normative framework for evaluating and accrediting engineering programs in Canada. The GAs is developed in consultation with industry professionals, ensuring that they are relevant to the needs of the engineering profession.	The GAs may not capture all the knowledge, skills, and abilities that are important for a particular engineering discipline or specialization due to the high level of abstraction.
Standard regulation (CDIO)	The Conceive-Design-Implement-Operate (CDIO) proposes a standardized framework for engineering education, which can facilitate the development of high-quality, globally relevant engineering programs. The syllabus includes a set of learning outcomes and core competences that are intended to prepare students for engineering practice, and provides guidance on curriculum design, teaching methods, and assessment.	Overly prescriptive that may limit the ability of institutions to develop innovative or context-specific engineering programs.
Accreditation body (ABET)	The Accreditation Board for Engineering and Technology (ABET) is a non-profit organization that accredits programs in engineering, computing, and technology disciplines in the United States. The organization sets standards for programs based on input from industry professionals, educators, and government officials. Established standards for the education and training of engineers may provide a starting point for the definition of engineer profiles.	Potential gaps in the standard that fail to capture all the skills and abilities required for a modern professional engineer and limited focus to United States.
Accreditation body (ENAEE)	The European Network for Accreditation of Engineering Education (ENAEE) is a non-profit organization that accredits engineering programs in Europe. It promotes quality engineering education across Europe and beyond, so that engineering graduates are fully equipped to tackle the issues and rigor that is demanded by modern engineering projects. ENAEE accreditation can be a useful tool for ensuring quality in engineering education and defining engineer professional profiles.	There may be variation in the standards applied across different programs and institutions, which can lead to inconsistency in the definition of professional engineering profile.

Semantic Indexing (LSI) (Deerwester et al., 1990), probabilistic LSI (Hofmann, 1999), and Latent Dirichlet Allocation (LDA) (Blei and Ng, 2003) have been widely adopted in academia and beyond for automating tasks manually inefficient due to the vast amounts of data involved (e.g., indexing, searching, summarizing, and extracting hidden topics) within a corpus (Chen et al., 2016; Vayansky and Kumar, 2020).

Topic modelling techniques hold great promise and require no additional training data, making them widely applicable in practical settings (Blei and Ng, 2003). Applications for topic modelling are numerous, from materials sciences engineering (Rani and Kumar, 2021), to software engineering (Chen et al., 2016; Silva et al., 2021), and research topics in engineering education (Johri et al., 2011). Many tools and analysis are available, including professional and open-source software, making topic modelling accessible to researchers of all levels (Jelodar et al., 2019).

A more recent development of topic model, which can potentially improve the interpretability of topics discovered, is to involve humans to iteratively refine the topics produced by an automated topic-modeling algorithm (Chuang et al., 2015). This approach is known as human-in-the-loop graphic interface. Interpretation can be boosted using the intertopic distance maps, which are adopted to visualize topics clusters as bubble graphs on a reduced dimensional (i.e., two-dimensional) space. The topics' projection on a reduced number of dimensions is based on multidimensional scaling (MDS) that aims to preserve the distances among topics (Sievert, 2014).

2.4. LDA for engineer archetype definition

LDA are a family of topic modelling techniques that automatically discover *latent* structure within an unstructured corpus of documents, using the statistical properties of its word frequencies (Blei and Ng, 2003; Jelodar et al., 2019). These models are branched off from the

subject area of generative probabilistic modeling in computational linguistics research (Liu et al., 2016). In this context, a *word* is considered as the basic unit of data, a *document* as a string of *n* words, and a *corpus* as a set of *m* documents that covers the dataset. The *vocabulary* is the collection of distinct words in the corpus, and a *topic* is a probability distribution across the vocabulary (Vayansky and Kumar, 2020). Specific words are expected to be seen more frequently in a document because a document is intuitively related to a specific topic. The topics discovered by topic modeling are semantic clusters created by words that are often used together (i.e., co-occur) in a document (Blei, 2012). For example, if a topic contains words, such as "human", "genome", "dna", "rna", "genetics", and "gene", this word cluster describes a topic related to "genetics" (Blei, 2012).

LDA algorithm attempts to decompose data into contributions from multiple latent topics that are shared by all the data, but to different extents. To elaborate, a typical topic model views each document as an unordered bag of words which occurs with different frequencies (Abdelrazek et al., 2023). It then "explains" the observed word frequencies in each document in terms of a suitably weighted mixture of topical word frequencies where the weights indicate the different proportions of topics that appear in the document (Guo et al., 2016; Jelodar et al., 2019). Practically LDA estimates the words in the text that are likely for a topic and topics that are likely for a particular corpus, using a fixed number of topics (i.e., k) decided a priori (Blei and Ng, 2003). The decision on the optimal value of k builds upon assessing perplexity and topic coherence metrics. The perplexity index assesses whether a statistical model fits well the dataset: the better is the estimation, the better the model and the lower the perplexity (Zhao et al., 2015; Maier et al., 2018).

LDA topic modelling has been used in past business research when attempting to profile emerging professional highly-tech related figures characterized by fast evolving and heterogeneous skills; hence, we deemed it an effective choice in terms of research design.

Showcases the value of LDA in job profiling research is provided to determine a universal Chief Digital Officers (CDO) archetype using as input a sample of 518 job postings scraped from LinkedIn and Indeed for CDO positions (Culasso et al., 2023). As highlighted by the authors, the need to draw a universal professional profile is urgent for those new roles that are emerging in response to digital transformation (Culasso et al., 2023).

Another application can be found in the field of Big Data job related positions. The authors highlighted a clear research gap regarding the formal definition of the most prominent Big Data jobs and of the required educational needs. They proposed a semi-automatic technique leveraged on a significant amount of online text data (i.e., 2.700 job posts which contained the keywords 'Big Data' in either the title or the description), obtained through web scraping, to generate an intelligible classification of job roles and keywords referring to skills, grouped by skill sets (De Mauro et al., 2018).

Other relevant works in the application of LDA for engineering mapping using online job advertisements can be found in (Gurcan and Cagiltay, 2019). Here the authors highlight how the education programs are now required to adapt themselves to up-to-date developments by identifying the knowledge domains and skill sets required for big data software engineering to meet the industrial needs and develop a taxonomy by mapping these competences (Gurcan and Cagiltay, 2019). The authors extracted 48 topics made by technical keywords and mapped them into 10 core competency areas (Gurcan and Cagiltay, 2019).

In order to generalize the previous highlighted applications, this work aims to define a framework for automatically defining the engineer archetype as a cluster of competences applying topic modelling to unstructured educational and occupational textual description using the well-known Gensim model (ldamodel; Řehůřek and Sojka, 2010). This work has the merit to push further the effort of previous works in this direction, providing an overarching understanding of what constitutes the contemporary professions of engineers in organizations, helping job marketplace to search for better recruitments and educators in developing human capital towards desired directions.

3. Method

As illustrated in Fig. 3, we propose a methodology based on text mining (light blue area) to automatically extract keywords that support the definition of an engineer archetype. Subsequently, the quality of the generated archetype is assessed manually by experts, and the input data and parameters are iteratively modified (light green area).

From left to right, the first crucial activity in green involves the manual definition of the domain of the archetype extraction. This activity sets limits on the research space and helps avoid scope creep of the professional figure under investigation.

The subsequent activities in blue are fully automated and form the core of the text mining process. They involve the preparation of a dataset comprising of heterogeneous textual descriptions from educational and occupational sources, according to the investigated domain. The dataset is preprocessed using NLP and the archetype is defined using LDA. These steps lead to the automatic discovery of latent topics that define the archetype as clusters of technical keywords.

The final group of green activities are manually performed by domain experts, who interpret the automatic results to provide meaningful labels to the extracted (unlabeled) topics. If the assessed archetype does not meet the expected outcome, the input of the automatic text mining process is iteratively updated manually. This involves modifying the initial premises, data source list, data preprocessing parameters, and LDA settings until satisfactory results are achieved.

3.1. Domain definition

The definition of domain (or scope) is the fundamental premise of the entire pipeline. It is crucial to correctly define the boundaries of the archetype (i.e., specific engineering field, geographical area, and time frame) and retrieve the propter educational and occupational data sources according. For those readers interested in defining a general engineer archetype, all the engineering fields could be addressed together. However, as a rule to be accounted for in these cases, the more the nested disciplines, the less the final output will be accurate and specific. A specific geographic area must be defined accordingly to the domain as well. A certain field could be differently studied (education) and applied (occupation) in different geographic or technical areas. Again, nested geographic areas could be considered, till the entire worldwide area, with the same effect previously highlighted: the more the generalization, the more the compensation factors.

As depicted in the top-left corner of Fig. 3, a potential output for the professional profile domain definition activity, and input for the subsequent automated text mining process for archetype generation activities, is a domain statement that contains the field, geographic region, time frame, educational data sources, and occupational data sources that should be considered when extracting textual descriptions of the professional figure in question.

As for educational-related text corpus the most reliable dataset can be directly retrieved by universities official websites. Based on the specific scope, a set of universities must be accounted for. The user can select specific universities based on specific needs or consider the top-rated ones as most representative of a specific area and field of study. In this latter case, the Times Higher Education (THE) World University Rankings, in partnership with Elsevier, includes almost 1400 universities across 92 countries, standing as the largest and most diverse university rankings ever to date (World University Rankings, 2021). It allows filtering results by different parameters such as geographic area, subject, and many others. Independently by the way the universities are chosen, the user must extract only a text corpus describing the skills, which the engineer will learn during the educational path. Courses description, syllabi, or Intended Learning Outcomes (ILOs) have a good

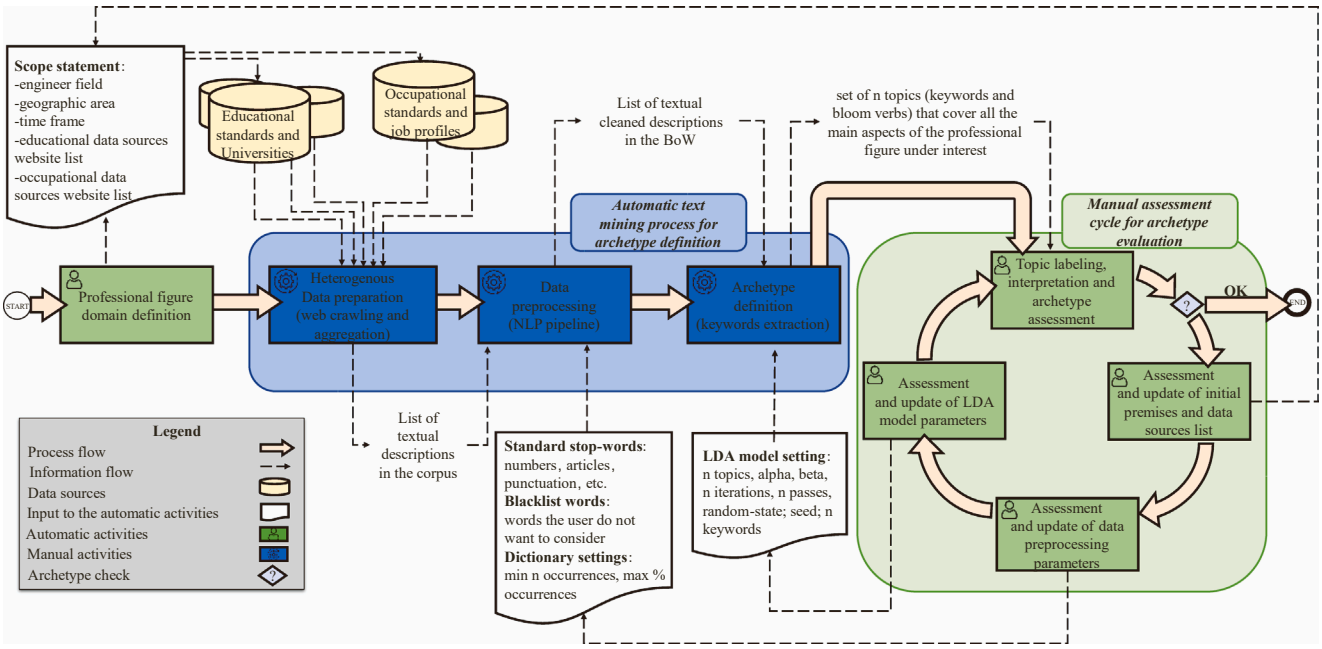

Fig. 3. Flow chart of the proposed methodology for the automatic extraction of archetypes.

chance to store the desired information. Another structured source of data for educational material are online libraries (Education.com).

Occupational text data can be effectively retrieved querying online job placement platforms (e.g., LinkedIn), allowing filtering the search based on specific job position (i.e., engineer field) and geographic area. These offer textual descriptions about open positions which by-design embed useful information about skills in the specific domain from an occupational perspective.

3.2. Data preparation

Once the domain is clear, a set of domain-specific textual descriptions of the professional profile under interest must be extracted and aggregated to prepare the input corpus file. Both occupational and educational data should be considered as the input to have a comprehensive spectrum of information from industry and education.

The information retrieval can be performed by using automatic web scraping or web crawling software able to extract the description content from the webpage list provided in the scope statement. From a structural perspective the heterogeneity of the data source requires careful attention in the extraction and listing as textual description. Websites have different HTTP protocols and copyright policies. For this reason, custom rules for the automatic crawling must be defined for each given application.

As shown in Fig. 3, the output information from the data preparation activity is a list of textual description items (i.e., "document") aggregated in a file formally named as "corpus", where each line corresponds to a course/job description. Each course description should have at least 100/200 words. The order of the rows is not constrained. Case letters, numbers, and lists are not constrained. No specific format is required. The raw output file can be an excel file or a simple txt.

3.3. Data preprocessing

As shown in Fig. 3, once the initial list of textual descriptions has been generated, a preprocessing step is required to clean the textual data and remove meaningless words (i.e., noise). For this task predefined vectors of stop word can be used (Chiche and Yitagesu, 2022). These essentially remove the Part of Speech (POS) which are out of the interest for semantic interpretation (e.g., punctuation, numbers, articles, adverbs, non-alphabetic character etc.) (Chiche and Yitagesu, 2022). Moreover, a customized blacklist of words which are not semantically interesting but recurrent in retrieved text descriptions can be removed as well (see case study in Section 4 as example). This list of words is defined manually and refined after checking the results of Section 3.4.

According to the preprocessing step, for each row of the input file (i.e., document or course/job descriptions), words are lemmatized (i.e., converting the words to their root form) and converted to lowercase, removing any words that match the blacklist or stop words list. As an additional step each word is mapped into a dictionary object via a unique ID and specific filters to the dictionary are applied. The information is a list of pre-processed tokenized documents, stored in the dictionary, and used to create a Bag of Words (BoW) representation by mapping the frequency of each word in the document to its corresponding index in the dictionary. Namely, words that appear in the vocabulary less than a given number of documents (i.e., courses/job description) are filtered out. Words that appear in more than a given percentage of the total corpus (i.e., all the courses/job description) are also removed as default. These two parameters are crucial and can be defined as follow:

- *Minimum document frequency*: The minimum number of documents a token must appear in to be included in the vocabulary.
- *Maximum document frequency*: The maximum percentage of documents a token can appear in to be included in the vocabulary.

This step is a common technique used in NLP to convert raw text data into a numerical format that can be used for further analysis, such as topic modeling, clustering, or classification.

3.4. Archetype generation

After filtrating the vocabulary, the BoW is passed to the last automatic activity: topic modelling. The LDA algorithm is applied to discover latent topics that are present in the corpus. In the implemented tool, the Gensim model is used (ldamodel; Řehůřek and Sojka, 2010). Among LDA model from Gensim we briefly recap the main important hyperparameters:

- *Number of topics*: The number of topics is one of the most important parameters. It determines the granularity of the topics generated by the model. Choosing an appropriate number of topics depends on the size of the corpus, the nature of the documents, and the goals of the analysis.
- *Alpha and Beta*: control the sparsity of the document-topic and topic-word distributions, respectively. However, if the corpus is small or the topics are expected to be very distinct, it is possible to increase alpha to encourage sparser document-topic distributions. Similarly, if the vocabulary is large, increasing beta encourages sparser topic-word distributions.
- *Passes*: controls the number of times the model is trained on the corpus. Increasing the number of passes can improve the quality of the topics generated, but it also increases the training time.
- *Iterations*: controls the number of iterations to be run for each document during training. Each document is processed multiple times to update the topic assignments and the topic-word and document-topic distributions.
- *Random state and random seed*: control the randomness of the initialization of the topic-word and document-topic distributions, which can affect the results of the model. By setting the random_state parameter to a fixed value, it is possible to ensure that the model is initialized with the same random seed every time it is run, which can lead to reproducible results.

The final input requested from the algorithm is a list of bloom verbs, which are a set of words used to describe cognitive processes and learning outcomes in education, and their presence in learning or occupational descriptions can be indicative of the level of knowledge targeted for the archetype (Maffei et al., 2022). These verbs are thus highlighted within the extracted keywords, which include both content and verb keywords. The keywords are ordered in decrescent order according to the likelihood to belong to the given topic according to the LDA approach. The number of selected keywords, excluding the non-bloom non-infinite verbs, is not native parameter for the LDA but can be included in the algorithm.

The output of the automatic text mining process for archetype definition is a set of n topics (keywords and bloom verbs) that cover all the main aspects of the professional figure under interest.

3.5. Experts' interpretation and topic labeling

After the automatic text mining process for archetype definition, the manual assessment cycle for archetype evaluation requires a first activity in which users assess the extracted topics as well as label the semantic of the topics behind. The pyLDAvis library provides an interactive visualization of topic models, which can be used to interpret the topics identified by the model (pyLDAvis). The graphic interface provides the list of words for each identified topic along with a bar chart that shows the most frequent terms in the selected topic. Such words represent and describe the topic contents and verbs. The authors assign a label to each topic based on the list of the most frequent terms (Boffa and Maffei, 2021).

During the topic labelling and interpretation of the extracted archetype, the expert assesses the quality of the outcome and if required, performs several feedback activities to adjust the input information of the automatic process. In more detail, the first assessment activity concerns the evaluation of the domain definition and the input data source. In some cases, misaligned information can be retrieved or too few documents can be considered. In this case, the scope statement and the data sources must be updated (e.g., defining new web scaping rules or changing the target websites). Secondly, preprocessing parameters must be assessed in accordance with the outcome alignment with the expectation. Refinement of the preprocessing parameters, stop words and black words lists must be tailored based on the obtained outcome and it could request several cycles of tuning.

Finally, the topic modelling algorithm requires specific parameters to be optimized, including the number of topics, alpha, beta, passes, iterations, random state, seed, and the number of keywords. An iterative approach to selecting appropriate parameters must be adopted. Specifically, an approach could be the systematic variation of different combinations of parameters, while keeping the others constant and checking the results. The objective of this iterative process is to identify the parameter set that would yield the most suitable model for each given application (Maier et al., 2018).

4. Case study

This section presents the experimentation and implementation of the proposed approach in an Industrial Engineering case study, considering the archetype defined in (Lupi et al., 2022) as benchmark. Sections 4.1–4.5 concern the implementation of the automatic steps defined in Section 3. A brief recap of the input data and output for the benchmark archetype is reported in Section 4.6. The developed Python code for archetype extraction is publicly available at GitHub: *https://github.com/francescolupi/ArchetypeAPP*.

4.1. Domain definition

The scope of the industrial engineering archetype is made by management, production and mechanical educational and occupational data sources adopted for the manual definition of the benchmark and used as same reference in this work (Lupi et al., 2022). In the current case, the autonomic definition of archetype requires a high volume of textual information. For this reason, courses within the European partners of the Manufacturing Education for a Sustainable fourth Industrial Revolution (MAESTRO) project (Mestro project) have been collected taking in consideration the same master programs accounted for the benchmark (i.e., management, mechanical, production and industrial engineering), but without removing similar or duplicated courses, for a total of more than 120 course descriptions.

4.2. Data preparation

In the preparation of the input file, the following master courses descriptions have been used: Management Engineering (Engineering and management) and Mechanical Engineering (Mechanical engineering) from the Department of Management and Production Engineering, Polytechnics of Turin (Italy); Production Engineering and Management (Production engineering and management), Industrial Management (Industrial management), Engineering Mechanics (Engineering mechanics) from KTH Department of Production Engineering, KTH Royal Institute of Technology (Sweden); Management Engineering (Management engineering) and Mechanical Engineering (Mechanical engineering) from the Department of Civil and Industrial Engineering, University of Pisa (Italy); Mechanical Engineering (Mechanical engineering) and Advanced Manufacturing and Engineering Management (Advanced manufacturing engineering and management) from the Loughborough University (United Kingdom).

Regarding the acquisition of occupational data, and despite the potential to adopt numerous job-related sources for information retrieval, we have exclusively relied on the ESCO database to adhere to the established benchmark case. However, the list of competences found in the ESCO database does not provide sufficient material for the LDA algorithm, which demands multiple words semantically arranged in coherent paragraphs that offer contextual information for each document. Considering this limitation, we have opted to consider solely the general description of the industrial engineering according to ESCO platform (ESCO - Industrial and production engineers).

A total of 127 entries (or documents) have been collected, each per row in a spreadsheet file. The input file has been prepared using web crawling whenever feasible, although an initially manual approach

proved necessary due to the highly different web formats found across various websites.

4.3. Data preprocessing

According to Sections 3.3 and 3.4, Table 2 summarizes the main adopted hyperparameters and input for the developed tool for data preprocessing and archetype generation. These parameters set can be used as default configuration for the first-time initialization of the algorithm, however it is recommended to experiment with different values and evaluate the results over several runs.

As the data preprocessing parameters, POS tagger spaCy is used (spaCy). Adjective (ADJ), Adverb (ADV), Auxiliary Verb (AUX), Interjection (INTJ), Pronoun (PRON), Coordinating Conjunction (CCONJ), Punctuation Mark (PUNCT), Particle (PART), Determiner (DET), Preposition or Postposition (ADP), Space (SPACE), Numeral (NUM) and Symbol (SYM) are considered as words to be removed.

The customized blacklist of words is made by those words that are highly likely to appear in many topics and do not provide specific information on the topic itself. Words such as "student" have been added to the blacklist by-design in the initial phase, other words have been identified after running the algorithm and checking the top 30 keywords. Additional words can be added if they possess no meaning in the topic's context.

The minimum document frequency and maximum document frequency has been selected by slightly varying the default settings.

4.4. Archetype generation

Regarding the generation parameters for archetypes, the first of interest for LDA setting is the optimal number of topics. One way to determine the optimal number of topics is to run the model with different numbers of topics and evaluate the coherence scores of the topics generated (perplexity measure). To comparing with the benchmark, 6 topics have been adopted.

The default values of alpha and beta are 1.0/num_topics, which work well in many cases. For passes, a good starting point is to use two or three. Both random state and random seed have been fixed to have repeatable results. The list of Bloom verbs has been generated considering the literature (Maffei et al., 2022; Krathwohl, 2010) and adding synonyms. A total of 100 keywords have been selected to be exported for each identified topic.

Together with the preprocessing parameters, the archetype generation parameters can be found in Table 2.

4.5. Experts' interpretation and topic labeling

The previous automatic text mining steps provide a list of keywords grouped into unlabeled topics (or clusters) as output. Manual interpretation of these topics is necessary to provide informative content to each topic based on the semantics of the keywords and to assign them meaningful labels. This interpretation relies on expertise. Since the visualization can help to identify patterns and relationships between topics, as well as to explore the most important terms associated with each topic, the interpretation and labeling of the six topics (T1-T6) generated for the current case study have been performed by the experts using scatter plots (Fig. 4–Fig. 9), which provide a graphical view of an industrial engineer within the geographic consortium of universities considered (MAESTRO).

The distribution of topics in two-dimensional space is provided using a technique called multidimensional scaling (MDS) to represent the similarity between topics. The size of each circle represents the prevalence of the topic in the corpus (i.e., marginal topic distribution), and the closeness of circles indicates that the topics share many common terms (i.e., semantic relation). The top 30 most salient terms for each topic, ranked by their frequency in the corpus, are also shown. The relative width of each term in the chart in red indicates its frequency within the topic (i.e., relevance) and the gray bars represent the overall frequency of the term in the corpus (i.e., saliency). The relevance of a term to a topic is a measure of how strongly that term is associated with the topic, based on its conditional probability and its overall frequency in the corpus. In the following, the six topics are visually depicted (red highlighted circles) from Fig. 4 to Fig. 9 and the main results discussed.

Fig. 4 (T1): Topic1 contains a group of keywords related to management and business. This cluster highlights the importance of various concepts such as strategy, control, and decision-making that an industrial engineer should possess. It also includes terms related to production, investment, and capital, emphasizing the financial aspect of managing a business. Planning is another central concept within this cluster, with words such as "plan," "planning," and "budgeting"

Table 2
The selected hyperparameter and tool (*https://github.com/francescolupi/ArchetypeAPP*) input for the case study.

Parameter	Description
Preprocessing POS vectors	['ADJ', 'ADV', 'AUX', 'INTJ', 'PRON', 'CCONJ', 'PUNCT', 'PART', 'DET', 'ADP', 'SPACE', 'NUM', 'SYM']
Customized blacklist	{'exercise', 'exercises', 'hour', 'hours', 'analysis', 'system', 'systems', 'student', 'students', 'expected', 'knowledges', 'method', 'methodology', 'methodologies', 'methods', 'problem', 'problems', 'model', 'models', 'modelling', 'project', 'projects', 'tutorial', 'tutorials', 'course', 'courses'}
Minimum document frequency	3
Maximum document frequency	90%
Number of topics	6
Alpha	1.0/Number of topics
Beta	1.0/ Number of topics
Passes	10
Iterations	Default
Random state	5
Random seed	5
Bloom verbs	["understand", "define", "identify", "describe", "label", "list", "name", "state", "match", "recognize", "select", "examine", "locate", "memorize", "quote", "recall", "reproduce", "tabulate", "tell", "copy", "discover", "duplicate", "enumerate", "listen", "observe", "omit", "read", "recite", "record", "repeat", "retell", "visualize", "explain", "interpret", "paraphrase", "summarize", "classify", "compare", "differentiate", "discuss", "distinguish", "extend", "predict", "associate", "contrast", "convert", "demonstrate", "estimate", "express", "infer", "relate", "restate", "translate", "ask", "cite", "discover", "generalize", "group", "illustrate", "judge", "observe", "order", "report", "represent", "research", "review", "rewrite", "show", "trace", "solve", "apply", "modify", "use", "calculate", "change", "choose", "discover", "relate", "sketch", "complete", "construct", "interpret", "manipulate", "paint", "prepare", "teach", "act", "compute", "list", "practice", "simulate", "write", "analyze", "classify", "contrast", "infer", "select", "categorize", "connect", "differentiate", "estimate", "evaluate", "focus", "organize", "plan", "question", "test", "reframe", "criticize", "appraise", "support", "decide", "recommend", "assess", "convince", "defend", "grade", "predict", "select", "argue", "conclude", "critique", "debate", "justify", "persuade", "weigh", "design", "compose", "plan", "combine", "formulate", "invent", "substitute", "compile", "develop", "integrate", "modify", "prepare", "rearrange", "adapt", "arrange", "collaborate", "facilitate", "make", "manage", "propose", "solve", "support", "test", "validate"]
Number of keywords	100

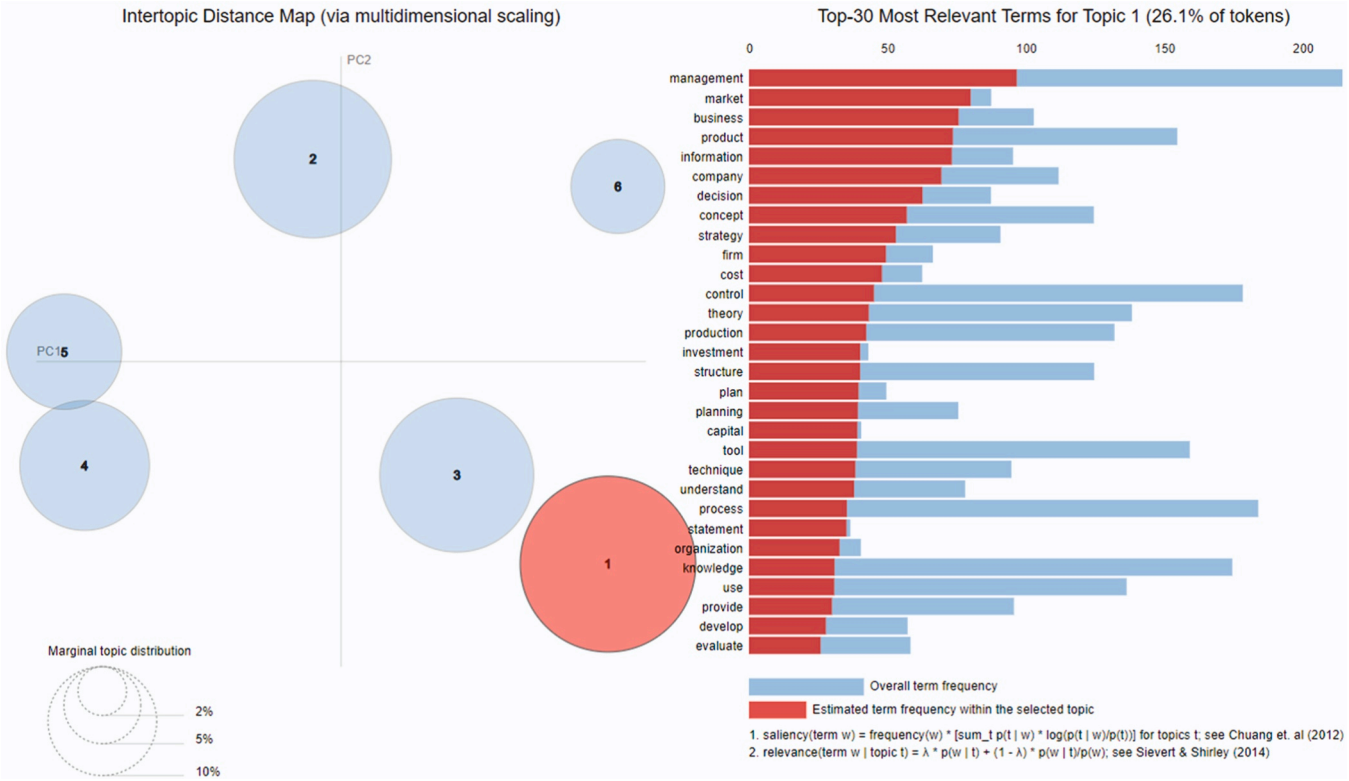

Fig. 4. The Topic1 (T1) is labelled as *Management, Finance, Economics and Business strategies*.

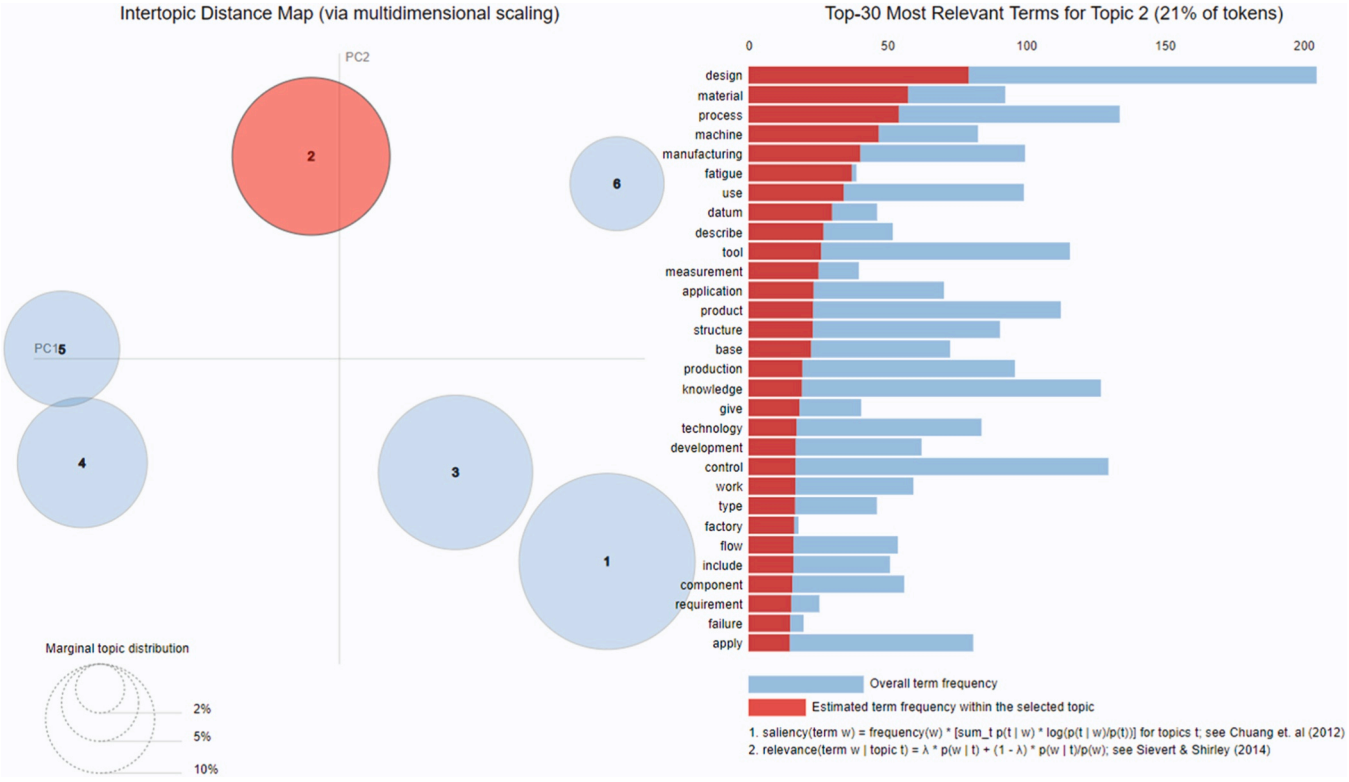

Fig. 5. The Topic2 (T2) is labelled as *Design of product and systems for quality*.

indicating the need for a well-thought-out strategy for success. The keywords also touch on the importance of understanding various processes, as well as the organization and management of resources for an industrial engineer. This topic also highlights the significance of technology and its role in the manufacturing industry. Additionally, terms such as "innovation," "growth," and "change" emphasize the need for

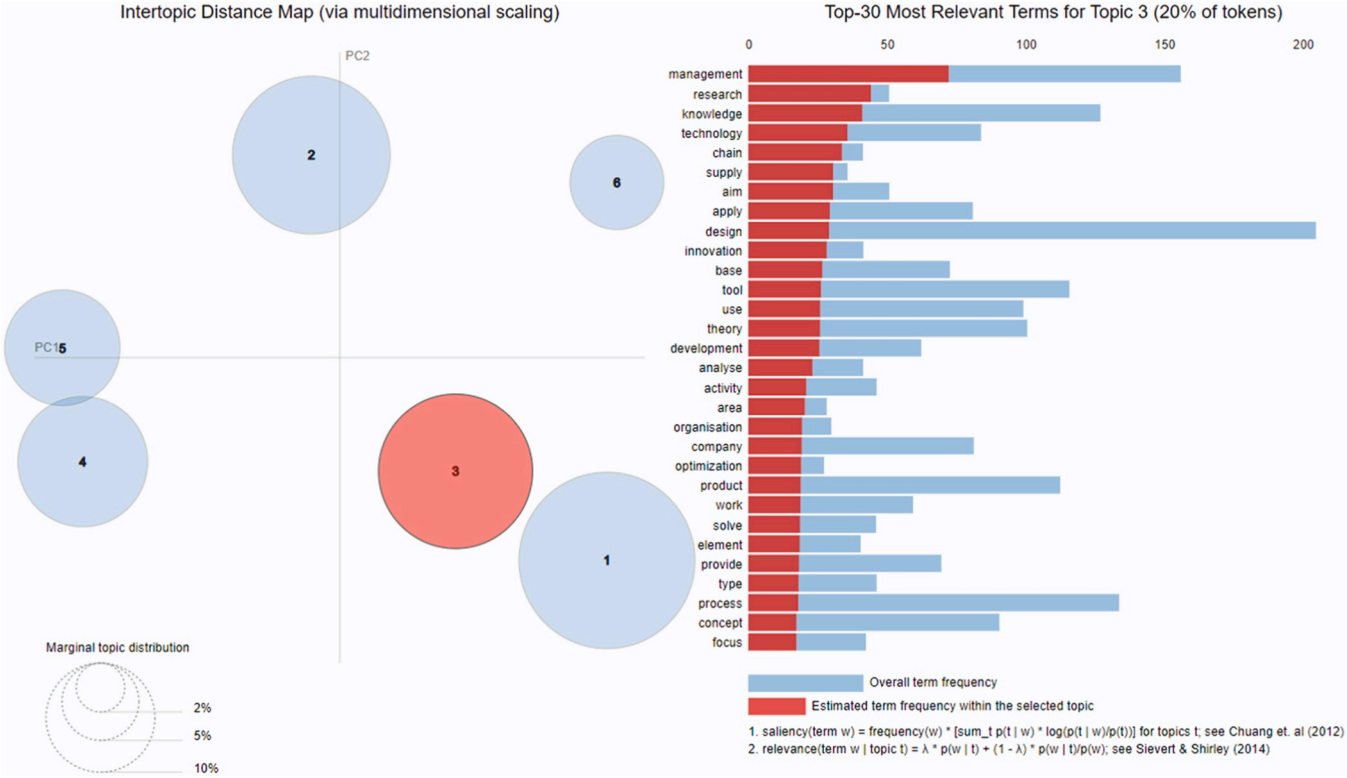

Fig. 6. The Topic3 (T3) is labelled as *Innovation and supply chain management*.

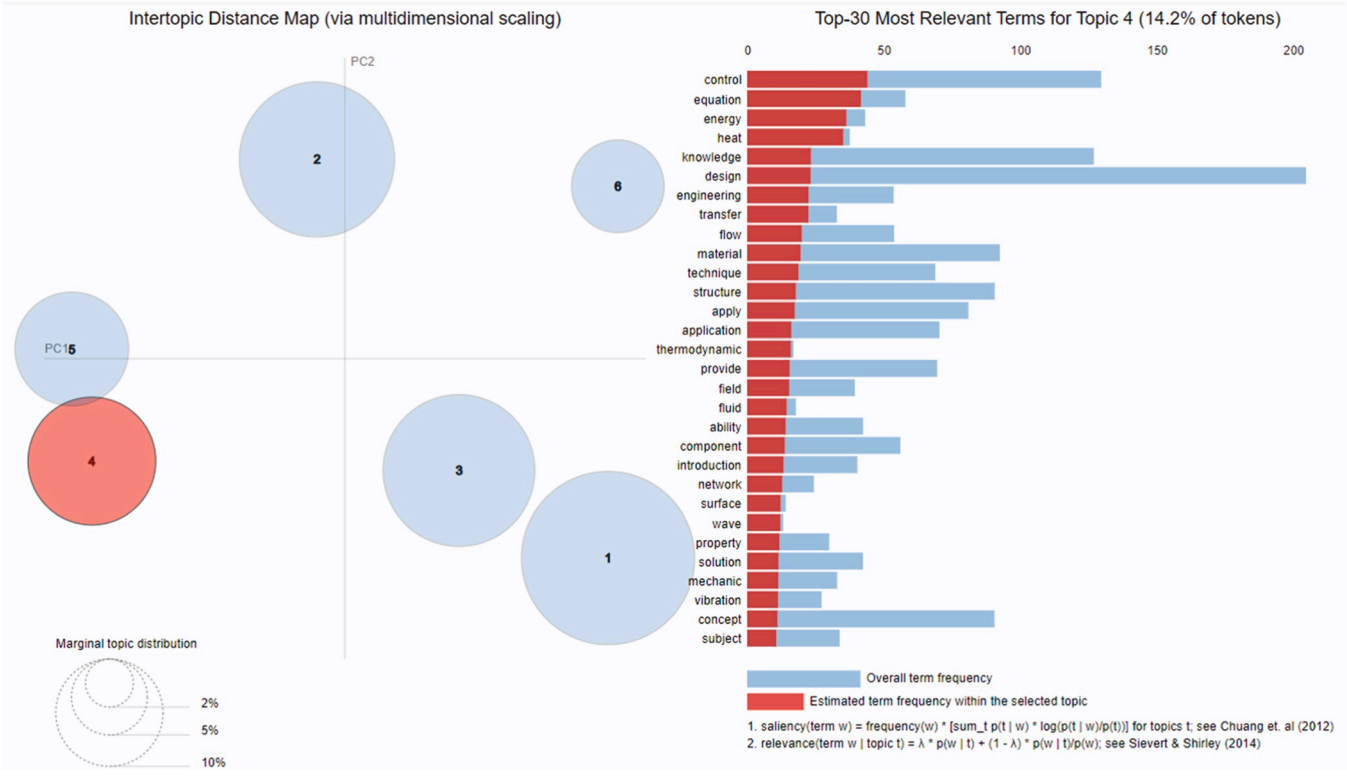

Fig. 7. The Topic4 (T4) is labelled as *Applied Thermodinamics and Mechanics*.

adaptation and evolution within a business. Finally, keywords such as "competency," "performance," and "functioning" indicate the importance of personal and team-based skills in the successful management of a business. Overall, this cluster emphasizes the importance of effective management, strategic planning, and innovation in the success of a business. This cluster has been labelled as **T1: Management, Finance, Economics and Business strategies.**

Fig. 5 (T2): Based on the keywords, this topic seems to be focused on

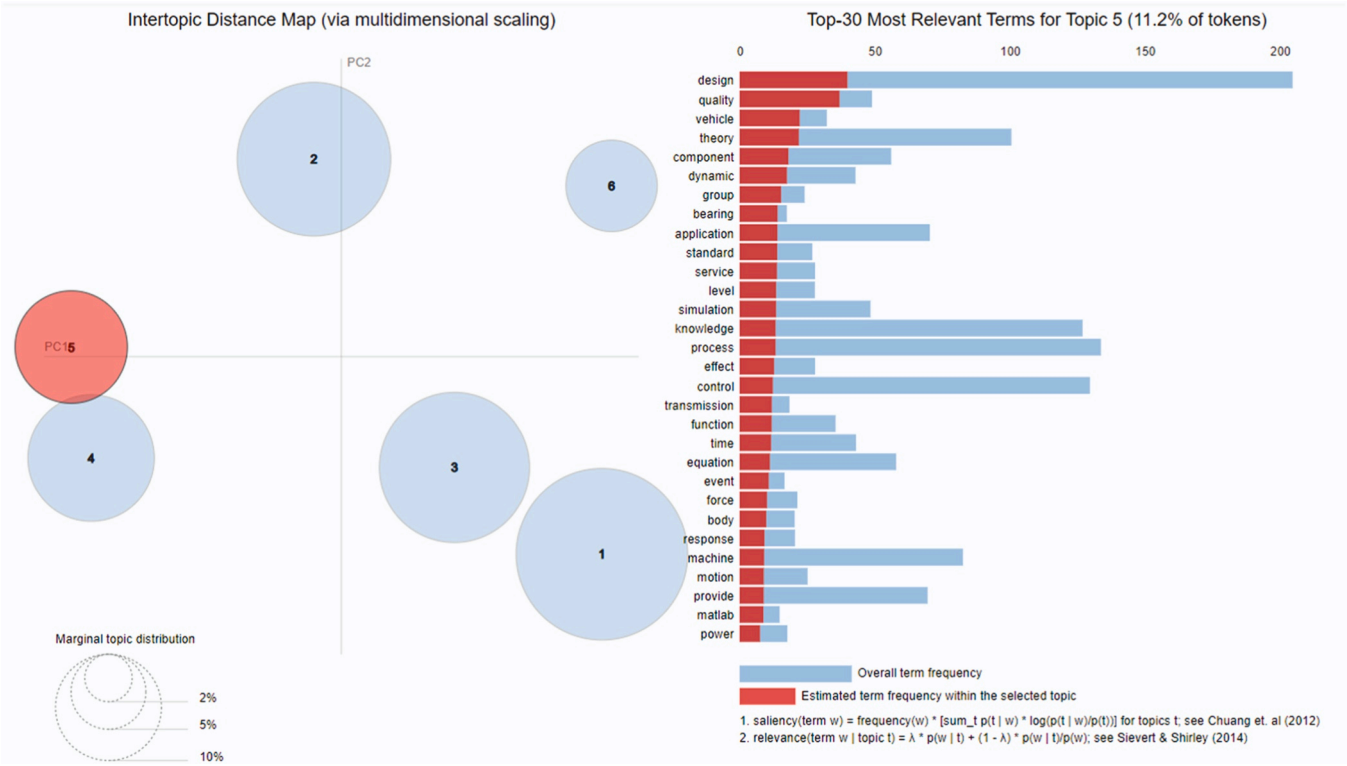

Fig. 8. The Topic5 (T5) is labelled as *Manufacturing and design IT tools*.

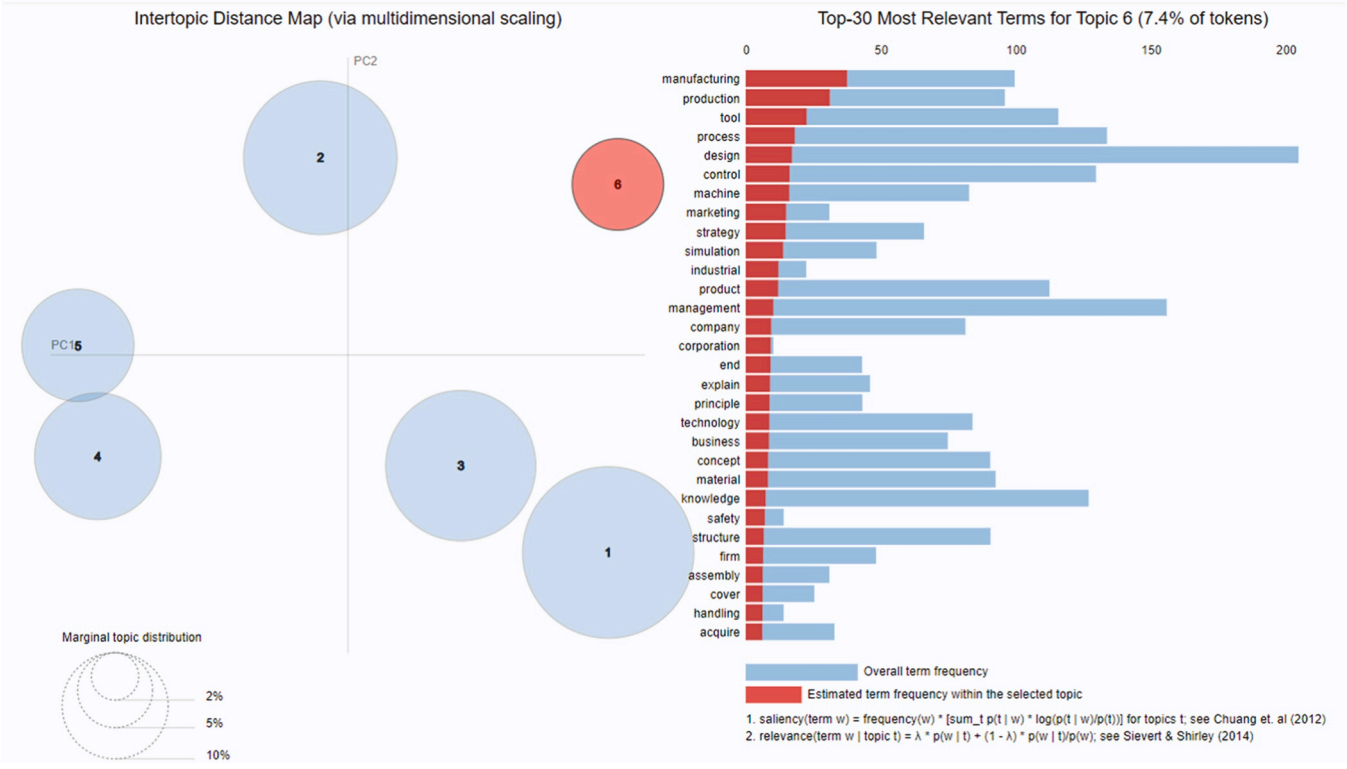

Fig. 9. The Topic6 (T6) is labelled as *Manufacturing, production planning and industrial automation*.

various aspects related to product design and manufacturing, including materials, processes, tools, and measurement. It also includes competences related to quality control, failure analysis, maintenance, and reliability. The presence of words like "simulation," "computation," and "modeling" suggest that this cluster may also relate to using computer-based tools to aid in the design and manufacturing process. Other keywords such as "datum" and "metrology" suggest that measurement and data analysis are important aspects of this cluster as well. Summarizing,

Table 3
The industrial engineer archetype as a set of 6 topics (T1-T6) automatically extracted using the proposed tool. The first column is the topic interpretation, while the second column presents the first 100 keywords excluding infinite verbs. Bloom verbs are enclosed by * and reported in the last column for clarity. Note that the words are ordered in decrescent order respect the relative importance to the specific topic.

Experts-based Topic Interpretation	Keywords ordered based on their weight	Bloom keywords verb
T1: Management, Finance, Economics and Business strategies	['management', 'business', 'product', 'information', 'decision', 'concept', 'strategy', 'firm', 'control', 'theory', 'production', 'investment', '*plan* ', '*planning* ', 'capital', 'tool', 'technique', '*understand* ', 'process', 'statement', 'organization', 'knowledge', '*use* ', 'provide', '*develop* ', '*evaluate* ', 'value', 'include', 'skill', 'financing', 'technology', 'accounting', 'manufacturing', 'activity', 'fundamental', 'follow', '*design* ', 'base', 'development', 'marketing', 'resource', '*relate* ', 'policy', 'need', 'industry', 'introduction', 'competition', 'principle', '*analyze* ', '*support* ', 'rate', 'task', 'case', 'role', '*apply* ', 'flow', 'implementation', 'present', 'demand', 'definition', 'deal', 'budgeting', 'work', 'portfolio', 'divide', 'opportunity', '*manage* ', 'innovation', '*understanding* ', 'ability', 'performance', 'framework', '*illustrate* ', 'know', 'competency', 'income', 'master', 'learn', 'growth', 'manager', 'functioning', 'idea', 'operation', '*change* ', '*discuss* ', '*focus* ']	['plan', 'planning', 'understand', 'use', 'develop', 'evaluate', 'design', 'relate', 'analyze', 'support', 'apply', 'manage', 'understanding', 'illustrate', 'change', 'discuss', 'focus']
T2: Design of products and systems for products and quality	['*design* ', 'material', 'process', 'manufacturing', '*use* ', 'datum', '*describe* ', 'tool', 'measurement', 'application', 'product', 'base', 'production', 'knowledge', 'give', 'technology', 'development', 'control', 'work', 'factory', 'flow', 'include', 'component', 'requirement', 'failure', '*apply* ', '*planning* ', 'theory', 'mechanic', 'maintenance', '*identify* ', 'simulation', '*explain* ', 'subject', 'plant', 'element', 'information', 'analyse', 'algorithm', 'life', 'pass', 'learn', '*estimate* ', '*develop* ', 'parameter', 'need', 'vibration', 'laboratory', '*understand* ', 'computer', 'deal', 'result', 'aspect', 'lab', '*focus* ', 'solution', 'aim', 'technique', 'layout', 'engineering', '*testing* ', 'behaviour', '*grade* ', '*complete* ', '*support* ', 'cam', '*solve* ', 'computation', 'processing', 'perform', 'standard', 'reliability', 'metrology', 'account', 'capability', 'response', 'modeling', 'calculation', 'property', 'quality', '*choose* ', 'issue', 'damage', 'crack', 'identification', 'manufacture']	['design', 'use', 'describe', 'apply', 'planning', 'identify', 'explain', 'estimate', 'develop', 'understand', 'focus', 'testing', 'grade', 'complete', 'support', 'solve', 'choose']
T3: Quality and supply chain management	['management', '*research* ', 'knowledge', 'technology', 'chain', 'supply', 'aim', '*apply* ', '*design* ', 'innovation', 'base', 'tool', '*use* ', 'theory', 'development', 'analyse', 'activity', 'area', 'organisation', 'optimization', 'product', 'work', '*solve* ', 'element', 'provide', 'process', 'concept', '*focus* ', '*change* ', '*formulate* ', 'engineering', 'industry', 'assembly', 'fem', 'science', 'perspective', '*describe* ', '*discuss* ', 'role', '*explain* ', 'field', '*evaluate* ', 'production', 'solution', 'case', 'ability', 'programming', 'data', 'result', 'strategy', 'sustainability', 'business', 'operation', '*analyze* ', 'give', 'create', 'carry', 'aspect', '*relate* ', 'technique', 'reflect', 'formulation', 'decision', 'environment', '*understand* ', '*make* ', 'need', 'pass', 'application', 'participant', 'complexity', 'control', '*choose* ', 'value', 'demand', 'cover', 'matrix', 'sector', 'algorithm', '*planning* ', '*connect* ', 'skill', 'industrial', 'present', '*develop* ', 'plane', 'degree', 'constraint', 'dynamic']	['research', 'apply', 'design', 'use', 'solve', 'focus', 'change', 'formulate', 'describe', 'discuss', 'explain', 'evaluate', 'analyze', 'relate', 'understand', 'make', 'choose', 'planning', 'connect', 'develop']
T4: Applied Thermodinamics and Mechanics	['control', 'equation', 'energy', 'knowledge', '*design* ', 'engineering', 'transfer', 'flow', 'material', 'technique', '*apply* ', 'application', 'thermodynamic', 'provide', 'field', 'fluid', 'ability', 'component', 'introduction', 'property', 'solution', 'mechanic', 'vibration', 'concept', 'subject', 'principle', '*analyze* ', 'dynamic', '*solve* ', 'phenomenon', 'radiation', 'form', 'formulation', 'device', 'distribution', 'involve', '*state* ', 'vehicle', 'linear', 'case', 'present', '*order* ', '*discuss* ', 'tool', 'numerical', 'stability', '*use* ', 'set', '*understand* ', 'concern', 'characteristic', 'reliability', 'result', '*relate* ', 'momentum', 'aspect', 'law', 'interaction', 'work', 'gas', 'programming', 'automation', 'example', 'objective', 'approximation', 'conservation', 'acquire', 'module', 'transmission', 'behavior', 'pipe', 'convergence', 'performance', 'emphasis', 'calculation', 'learn', 'physics', 'robot', 'etc', 'modeling', 'theory']	['design', 'apply', 'analyze', 'solve', 'state', 'order', 'discuss', 'use', 'understand', 'relate']
T5: Manufacturing and design IT tools	['*design* ', 'quality', 'vehicle', 'theory', 'component', 'dynamic', '*group* ', 'bearing', 'application', 'standard', 'level', 'simulation', 'knowledge', 'process', 'control', 'transmission', 'equation', 'event', 'force', 'response', 'provide', 'matlab', 'criterion', '*relate* ', 'introduction', 'gear', 'capability', 'frequency', 'linear', '*apply* ', 'connection', 'selection', 'leadership', 'h', 'know', 'include', 'concept', 'base', 'etc', 'stiffness', 'verification', 'mean', 'implementation', 'behaviour', 'track', '*solve* ', 'tool', 'performance', 'classification', 'mode', 'matrix', 'regression', 'fundamental', 'sampling', '*test* ', 'bear', 'parameter', 'organisation', 'oscillation', 'kinematic', 'vibration', 'vector', '*explain* ', '*support* ', 'coefficient', 'series',	['design', 'group', 'relate', 'apply', 'solve', 'test', 'explain', 'support', 'use', 'state']

(continued on next page)

Table 3 (*continued*)

Experts-based Topic Interpretation	Keywords ordered based on their weight	Bloom keywords verb
T6: Manufacturing processes, production planning and automation	'euler', 'domain', 'technique', 'software', 'field', '*use* ', 'distribution', 'chain', 'correlation', 'trajectory', 'implement', '*state* ', 'point'] ['manufacturing', 'production', 'tool', 'process', '*design* ', 'control', 'marketing', 'strategy', 'simulation', 'industrial', 'product', 'management', 'corporation', '*explain* ', 'principle', 'technology', 'business', 'concept', 'material', 'knowledge', 'safety', 'firm', 'assembly', 'cover', 'handling', 'acquire', 'line', 'relationship', '*order* ', 'work', '*use* ', 'create', 'property', 'computer', 'datum', 'plc', 'data', '*focus* ', 'context', 'manufacture', 'plant', 'exchange', 'code', 'sensor', 'pillar', 'machining', 'operation', '*planning* ', 'consumer', 'lean', 'emphasis', 'workplace', 'integration', 'configuration', 'agency', 'environment', 'deployment', 'programming', 'preparation', 'shareholder', 'organisation', 'opportunity', 'consider', 'example', 'regulation', 'mechanism', 'difference', 'development', 'equipment', 'law', 'ce', 'machinery', 'introduction', 'manual', 'jit', 'perspective', 'role', 'wcm', 'flow', 'build', '*compare* ', 'sector', 'decision', 'module', 'provide', 'application', 'part', '*understand* ', 'kpi']	['design', 'explain', 'order', 'use', 'focus', 'planning', 'compare', 'understand']

this cluster appears to be related to the operational aspects of product design and manufacturing, including the use of tools and techniques to optimize manufacturing processes and ensure product quality. According to this interpretation, the topic has been labelled as **T2: Design of products and systems for quality**.

Fig. 6 (T3): Topic3 is quite close to the T1 but is more focused on the management and application of technology. The words "management", "technology", "supply chain", and "innovation" all appear prominently in this cluster, as well as words related to organizational optimization, product development, and sustainability. There is also a focus on using tools and techniques to analyze and solve problems, as well as understanding the complexity of different industries and sectors. In general, this cluster seems related to the practical application of research and technology in different settings, with an emphasis on management and optimization. Words related to decision-making, strategy, and planning are also present, suggesting a focus on achieving specific goals and outcomes. The cluster **T3** has been labelled as **Innovation and supply chain management.**

Fig. 7 (T4): This topic seems to be focused on competences related to control, energy transfer, and fluid mechanics, with a strong emphasis on applying knowledge to practical applications. Aspects related to mathematical modeling and simulation are also present, indicating a focus on using analytical tools to solve problems related to engineering and physics. There is also an emphasis on understanding the fundamental principles underlying various phenomena, such as thermodynamics, mechanics, and radiation. Programming and automation are also mentioned, suggesting that the cluster may relate to topics such as robotics and process control. Finally, there are some more general keywords related to reliability, performance, and physics, suggesting that this cluster may encompass a broad range of engineering and science disciplines. This topic has been named T4: **Applied Thermodinamics and Mechanics**.

Fig. 8 (T5): Based on the keywords in this topic, it seems to be focused on competences related to design and engineering, particularly in the manufacturing industry. Some specific concepts that appear to be important in this cluster include quality, component, dynamic, bearing, transmission, gear, stiffness, and performance. Other words like "simulation," "control," "parameter," and "verification" suggest a focus on testing and analysis to ensure high-quality designs. The presence of words like "leadership," "support," and "organisation" also suggests a focus on team collaboration and project management. T5: **Manufacturing and design IT tools**.

Fig. 9 (T6): Based on the keywords extracted, it seems that Topic6 may be related to manufacturing and production, specifically topics such as industrial design, process control, product management, technology, and business strategy. Other keywords in this cluster suggest a focus on manufacturing operations, including assembly, machining, operation, and lean principles. Other keywords suggest a focus on data and computer technology, including "datum," "data," "PLC," "code," and "computer." Overall, it seems that cluster 6 is focused on the operational and technical aspects of manufacturing and production, with an emphasis on industrial process automation. This topic has been interpreted and labelled as T6: **Manufacturing processes, production planning and industrial automation**.

As a result, the generated archetype is a list of technical (i.e., content) and bloom verbs (i.e., verb) keywords clustered into unlabeled topics which, after expert interpretation, are labeled and presented in Table 3. No context has been extracted for the sake of brevity and to maintain the archetype in a concise form. From Bloom verbs standpoint, as general but remarkable result, the 'design' verb has emerged as main verb for all the generated clusters of the industrial engineer (i.e., it is ranked in the first top verbs for each topic), confirming prominence of this high-level cognitive activity for the engineer archetype.

4.6. *The industrial engineering benchmark*

The input data pertains to information sourced from the course syllabi of university master degree programs within the MAESTRO project (Project description). In order to limit the manual effort required for the analysis of a large volume of data, 50 courses have been manually selected for use in the benchmark (Lupi et al., 2022). For a specific course, constraints have been posed in extracting a synthetic description focusing on what the students should be able to do at the end of the course considering the three pillars of the Intended Learning Outcome ILOs (i.e., acting verbs, contents, and contexts) (Maffei et al., 2022). Since these three pillars' presence is not mandatory, these descriptions were called "semi structured" ILOs (i.e., SS-ILOs).

As for the occupational side, the input information has been retrieved via the multilingual classification of ESCO (Esco). Given a specific job, the related skills can automatically be retrieved in the database. Compared to SS-ILOs, this information set presents much more synthetic competences that often embed verbs and contents of the competence itself.

The elements of the two input sets have been grouped, and clustered through a consensus-based approach by groups of academic experts. The main task of this activity has been the synthesis of the input list to a final set of 6 clusters of competences that cover different main topics (i.e., CL1. Manufacturing Processes; CL2. Structures, Machines, and Products Design; CL3. Production IT Tools Infrastructure; CL4. Manufacturing Automation and Robotics; CL5. Production Planning and Control; CL6.

Logistics and Supply Chain Management).

The respective clusters of competences (i.e., the same of topics in the current work) hence define the engineer archetype. The full description of the industrial engineer archetype can be found in (Lupi et al., 2022).

5. Discussion

The significance of this endeavor lies in kickstarting the automation of engineer archetype definition using text mining techniques for direct extraction of extensive text corpus from online sources such as public educational websites and job placement platforms. The current work boasts several advantages and benefits, including:

- Quick and reliable processing of text using trending open-source algorithms developed in the NLP domain.
- Quick and easy to update archetypes without the need to train the algorithm. University and industry are non-static entities, and archetypes should reflect these dynamics (e.g., the Industry 4.0 revolution in the recent years).
- Synthetic representation of a massive number of instances increases the accuracy of the output archetype with respect to smaller ones generated using smaller input datasets processable via a manual approach.
- Standardized archetype's definition via a common language between the interested parties (i.e., university and industry). This guarantees a robust procedure that delivers de-facto profiles, which enable the realignment of the educational and occupational frameworks.
- Broaden the application of this technique to any field of study, even further the application to only the engineering domain but also to any professional archetypes (e.g., physician, biologist, psychologist, designer).
- Online accessible tool for archetype definition at https://github.com/francescolupi/ArchetypeAPP.

The next Sections provide the comparison between the developed automatic approach as opposed to the manual method adopted as benchmark (Section 5.1), parameters tuning and setting (Section 5.2.) and discussion on the open challenges as well as possible future development (Section 5.3).

5.1. Comparison of manual expert and automatic NLP-based results

A high-level comparison has been conducted between the automatic and manual outcomes generated using the same input dataset to validate the proposed approach against the benchmark (Lupi et al., 2022). Specifically, the six topics (T1-T6) of the automatically defined archetype are compared to the six clusters of competences (CL1-CL6) manually defined in our previous work.

T2: Design of Products and Systems for Quality involves designing, analyzing, and testing products and systems, focusing on manufacturing, control theory, and quality assurance. Likewise, CL2: Structures, Machines, and Products Design involves researching, designing, and testing machines, mechanical installations, and components, materials analysis, preparation of technical documentation and experiments. Another similarity can be found in CL4: Manufacturing Automation and Robotics, which involves the application of control theory, and design in manufacturing systems. It includes selecting components, assembling machines, and programming systems for automation, material handling, and assembly processes.

T3: Quality and Supply Chain Management focuses on management, research, and technology within the supply chain. Emphasizes applying design, innovation, optimization, and problem-solving in the industry. Similarly, CL6: Logistics and Supply Chain Management involves applying mathematical models for demand anticipation, optimizing aggregated planning, inventory management, and resource exploitation. It includes monitoring supply flow, managing supply chain activities, and synchronizing supply with customer demand.

T5: Manufacturing and Design IT Tools focus on design, simulation, control, and application of industrial-specific software tools and applications. CL3: Production IT Tools Infrastructure involves practical skills in using CAE software for integrated manufacturing systems and focuses on evaluating, selecting, and designing optimal ICT solutions and PLM systems.

T6: Manufacturing Processes, Production Planning, and Automation involves control, simulation, assembly, and flow optimization, emphasizing production, technology, and manufacturing processes design for efficient production. Similarly, CL1: Manufacturing Processes includes designing conventional and non-conventional systems, optimizing assembly technology, simulating flow, and creating factory layouts using specialized software. Additional partial overlapping can be found in CL5: Production Planning and Control focuses on designing and optimizing production planning, maintenance processes, statistical monitoring, Health, Safety & Environment (HSE) systems, and improving production efficiency and costs. Other, similarities can be found with CL4: Manufacturing Automation and Robotics, which involves the application of robot modeling, control theory and automation.

The two topics that have a one-to-one or strong alignment with specific clusters from the benchmark are T1: Management, Finance, Economics, and Business Strategies and T4: Applied Thermodynamics and Mechanics. These two new recognized topics by the automatic algorithm are due to the tendency of the algorithm to maximize the interclassed dissimilarities and merging close topics to single ones, which may result in nested topics. In this regard, if certain emerged topics appear to be too generalist, it is possible to consider only a subset of competences that primarily belong to specific professional categories during the expert interpretation. This flexibility may allow to tailor the archetype to accurately represent the desired professional profile and ensure its relevance within a specific context.

From an initial comparison between the two industrial engineering archetypes, significant overlap is shown. Both the former (i.e., the manually generated benchmark) and the latter (i.e., the automatically generated one) consist of six categories (i.e., cluster of competences and topics). The overlap between several topics/clusters demonstrates the feasibility of automatically defining an engineer industrial archetype and providing a general framework and an online tool that can be used to generate other archetypes according to a specific domain. As an additional result, an entire topic regarding finance, economics and business model emerged in the current work, thanks to the processing of a higher volume of data, but still from the same source of the manually defined archetype.

From a "structural" perspective, the first archetype is more specifically related to the competences required for this type of professional profile, while the second archetype focuses on the technical keywords that are important in this field. This is due to the limit of topic extraction and automation of archetype definition, which makes it challenging to extract summary sentences of big input dataset and only restrict to keywords.

5.2. Parameters setting

The automatic approach of generating archetypes using the LDA topic extraction algorithm has advantages and limitations. One limitation is that the results of the LDA can be sensitive to the choice of hyperparameters, such as the number of topics and optimization may be required (e.g., perplexity and coherence metrics). The perplexity calculation is performed using a k-interval defined a priori (Blei and Ng, 2003). The result is a graph that shows the metric's value for each k and plots the trend line of the metric. The k-interval is restricted in correspondence with the trend line's elbow. The restricted k-interval will guarantee a low perplexity and exploitable numbers of topics. Nevertheless, the perplexity measure does not inspect the semantics. A topic-coherence metric is calculated over the restricted interval of k. The

metric checks within each topic if the included words are semantically related based on "co-document frequency of words" (Mimno et al., 2011). The higher the score, the higher the coherence and the topic interpretability. The optimal k chosen to run LDA corresponds to the higher coherence score value.

In the context of topic modeling, the length of the documents can impact the performance of the model. Shorter documents may be more difficult to model accurately because they contain fewer words and may not have enough context to support accurate topic identification. Conversely, very long documents may contain multiple topics or themes, making it more challenging for the model to identify the main topics (Albalawi et al., 2020). However, this can vary based on the specific use case and the quality of the data.

On the other side, the automatic approach allows for the processing of a larger and potentially enormous amount of data in a shorter time compared to the manual approach adopted in the benchmark. Furthermore, it can reduce bias and subjectivity in the archetype definition process.

5.3. Open challenges and future development

The current paper represents a step forward from our previous work, where both the data collection and archetype definition process were performed entirely manually by experts. In this paper, we demonstrate the feasibility of fully automated definition of archetypes using NLP and topic extraction modelling. However, the current data collection process still requires manual effort and is not completely based on web scraping.

Our ongoing work aims to leverage web scraping to fully automate the data collection process from diverse HTML sources. However, implementing a generalizable framework for web scraping is challenging due to the lack of standardization in the HTML structure of university and occupational websites. Future work should emphasize developing automated techniques for web scraping that can adapt to variations in HTML structure across different websites. This can be achieved by searching for specific HTML tags or keywords in the HTML content, and by creating a generalizable method for extracting course descriptions and job profile information from various websites (Kobayashi and Takeda, 2000).

If content (technical keyword) and Bloom verbs (bloom) are considered not enough for a complete description of the archetype, additional research for a summary extraction of the context of the topic may be required. A backward search in the most related documents to the extracted topics may be useful in searching for a context sentence (e.g., the most related contextual words to the topic under interest).

Using bigrams and trigrams, which are two or three consecutive words (e.g., "additive manufacturing," "computer-aided design"), in topic extraction can supply additional context and improve the accuracy of the results, rather than relying solely on individual words (i.e., unigrams).

Increasing the overlap between the educational and the occupational domains remains another ongoing challenge. Academia should focus on advancing emerging technologies at lower Technology Readiness Levels (TRLs), while industries continuously encounter new needs that may not be covered by academic institutions. These industry-specific challenges often remain unexplored by academics, leading to a gap at the intersection of educational research and practical occupational requirements. The light blue symmetric difference E $\triangle$ O in Fig. 2b represents two distinct areas of competences that must be carefully managed, without necessarily expecting complete overlap. These distinct competences between education and occupation should be preserved to a certain extent. Thus, a persistent gap exists at the frontier where educational research and occupational applications diverge. It is crucial to acknowledge and address this ongoing issue, and establish tools and criteria to assess this gap, enabling evaluation from both educational and occupational perspectives while comparing it against the archetypes, which is the first ingredient in this direction (i.e., reference for assesment feedback).

Other future developments efforts will focus on the practical application of our framework to define emerging engineering figures in fields characterized by non-standardized descriptions and ambiguous definitions. This includes areas like sustainability (Gutierrez-Bucheli et al., 2022), lean manufacturing (Chiera et al., 2021) and artificial intelligence (Yüksel et al., 2023). By addressing these challenges, our aim is to enhance the practicality of our research and make a valuable contribution to real-world industry settings. Through this work, we aim to bridge the gap between academia and industry by providing insights for the development of relevant and effective engineering figures using archetypes. Our commitment lies in making a significant impact in the engineering field through these endeavors.

6. Conclusion

Engineering archetypes have been defined here as clusters of competences and skills that have been identifined to simplify the communication, assessment, and comparison of technical topics required by different professionals in the engineering field. Their role is crucial in bridging the gaps between educational and occupational frameworks, promoting harmonization by aligning supply and demand. Additionally, they provide valuable insights for managers in decisions-making processes related to the selection and professional education. However, traditional manual methods for defining archetypes face limitations such as time-consuming, biased, and inconsistent approaches. These gaps include a lack of automatic methodologies, reliance on manual surveys, and limited coverage across programs and institutions. In response to these limitations, our contribution lies in presenting an abstract representation of a professional profile (i.e., archetype). The archetype principle was initially introduced in our previous paper, and now our objective is to automate the process of generating archetypes. Furthermore, we propose an iterative adoption and update of archetypes to continually assess and improve the alignment and harmonization of education and occupation over time. The results of our study demonstrate the feasibility of the proposed method in defining an industrial engineer archetype and provide an online tool that utilizes raw text data to generate archetypes within specific domains. The use of natural language processing techniques proves to be more objective and efficient in defining archetypes, leveraging large text corpora from accessible sources and employing open-source algorithms for accurate topic extraction. Furthermore, the proposed method aims to standardize and harmonize the definitions of professional engineering profiles, offering exciting opportunities for representative archetypes of the latest trends in the field. The findings of our study indicate a high degree of overlap between the expert-based and automated-based methods, validating the use of the automated methodology for future studies and improvements.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The authors are unable or have chosen not to specify which data has been used. The developed prototype software for archetype generation is publicly available online at https://github.com/francescolupi/ArchetypeAPP.

References

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A., 2023. Topic modeling algorithms and applications: a survey. Inf. Syst. 112, 102131 https://doi.org/10.1016/J.IS.2022.102131.

ABET | ABET Accreditation, (n.d.). ⟨https://www.abet.org/⟩ (accessed April 11, 2023).

About accreditation | Engineers Canada, (n.d.). ⟨https://engineerscanada.ca/accreditation/about-accreditation⟩ (accessed April 11, 2023).
Advanced Manufacturing Engineering and Management Degree | Postgraduate study | Loughborough University, (n.d.). ⟨https://www.lboro.ac.uk/study/postgraduate/masters-degrees/a-z/advanced-manufacturing-engineering-management/⟩ (accessed April 12, 2023).
Albalawi, R., Yeap, T.H., Benyoucef, M., 2020. Using topic modeling methods for short-text data: a comparative analysis. Front Artif. Intell. 3, 42. https://doi.org/10.3389/FRAI.2020.00042/BIBTEX.
A. Amado, P. Cortez, P. Rita, S.M.-E.R. on Management, undefined 2018, Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis, Elsevier. (n.d.). ⟨https://www.sciencedirect.com/science/article/pii/S2444883417300268⟩ (accessed April 11, 2023).
Blei, D., Ng, A., 2003. M.J.-J. of machine L. research, undefined 2003, Latent dirichlet allocation. Jmlr. Org. 3, 993–1022. ⟨https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com⟩ (accessed April 11, 2023).
Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55, 77–84. https://doi.org/10.1145/2133806.2133826.
Boffa, E., Maffei, A., 2021. Classification of Sustainable Business Models: A literature review and a map of their impact on the Sustainable Development Goals. FME Transactions 49 (4), 784–794.
Browse Educational Resources | Education.com, (n.d.). ⟨https://www.education.com/resources/?q=mechanical+engineer&cid=11.75⟩ (accessed April 12, 2023).
CDIO Syllabus | Worldwide CDIO Initiative, (n.d.). ⟨http://www.cdio.org/framework-benefits/cdio-syllabus⟩ (accessed April 11, 2023).
Chen, T.H., Thomas, S.W., Hassan, A.E., 2016. A survey on the use of topic models when mining software repositories. Empir. Softw. Eng. 21, 1843–1919. https://doi.org/10.1007/S10664-015-9402-8/TABLES/6.
Chiarello, F., Bonaccorsi, A., Fantoni, G., 2020. Technical sentiment analysis. measuring advantages and drawbacks of new products using social media. Comput. Ind. 123, 103299 https://doi.org/10.1016/J.COMPIND.2020.103299.
Chiarello, F., Trivelli, L., Bonaccorsi, A., Fantoni, G., 2018. Extracting and mapping industry 4.0 technologies using wikipedia. Comput. Ind. 100, 244–257. https://doi.org/10.1016/J.COMPIND.2018.04.006.
Chiche, A., Yitagesu, B., 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. J. Big Data 9, 1–25. https://doi.org/10.1186/S40537-022-00561-Y/FIGURES/5.
Chiera, M., Lupi, F., Rossi, A., Lanzetta, M., 2021. Lean maturity assessment in eto scenario. Applied Sciences 11 (9), 3833.
Chowdhary, K.R., 2020. Natural language processing. Fundam. Artif. Intell. 603–649. https://doi.org/10.1007/978-81-322-3972-7_19.
Chuang, J., Roberts, M.E., Stewart, B.M., Weiss, R., Tingley, D., Grimmer, J., Heer, J., 2015. TopicCheck: Interactive alignment for assessing topic model stability. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies*, pp. 175–184.
Courses for Engineering Mechanics | KTH | Sweden, (n.d.). ⟨https://www.kth.se/en/studies/master/engineering-mechanics/courses-engineering-mechanics-1.268705⟩ (accessed April 12, 2023).
Courses for Industrial Management | KTH | Sweden, (n.d.). ⟨https://www.kth.se/en/studies/master/industrial-management/courses-industrial-management-1.268639⟩ (accessed April 12, 2023).
Courses for Production Engineering and Management | KTH | Sweden, (n.d.). ⟨https://www.kth.se/en/studies/master/production-engineering-management/courses-production-engineering-management-1.268732⟩ (accessed April 12, 2023).
Culasso, F., Gavurova, B., Crocco, E., Giacosa, E., 2023. Empirical identification of the chief digital officer role: a latent Dirichlet allocation approach. J. Bus. Res 154, 113301. https://doi.org/10.1016/J.JBUSRES.2022.113301.
D.C. Davis, S.W. Beyerlein, I.T. Davis, Development and use of an engineer profile, ASEE Annual Conference and Exposition, Conference Proceedings. (2005) 4279–4292. https://doi.org/10.18260/1–2–14201.
De Mauro, A., Greco, M., Grimaldi, M., Ritala, P., 2018. Human resources for Big Data professions: a systematic classification of job roles and required skill sets. Inf. Process Manag 54, 807–817. https://doi.org/10.1016/J.IPM.2017.05.004.
Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41 (6), 391–407.
Doré, S., Cicek, J.S., Jamieson, M.V., Terriault, P., Belleau, C., Rodrigues, R.B., 2021. What is an engineer: study description and codeboook development. Proc. Can. Eng. Educ. Assoc. (CEEA). https://doi.org/10.24908/pceea.vi0.14850.
Esco, (n.d.). ⟨https://esco.ec.europa.eu/it/classification/occupation_main⟩ (accessed April 12, 2023).
Fareri, S., Fantoni, G., Chiarello, F., Coli, E., Binda, A., 2020. Estimating Industry 4.0 impact on job profiles and skills using text mining. Comput. Ind. 118 https://doi.org/10.1016/J.COMPIND.2020.103222.
Floyd, I.R., Jones, M.Cameron, Twidale, M.B., 2008. Resolving incommensurable debates: a preliminary identification of persona kinds, attributes, and characteristics. Artifact 2, 12–26. https://doi.org/10.1080/17493460802276836/ART.2.1.12_1.
J. Gainsburg, C. Rodriguez-Lluesma, D.E. Bailey, A "knowledge profile" of an engineering occupation: temporal patterns in the use of engineering knowledge, Https://Doi.Org/10.1080/19378629.2010.519773. 2 (2010) 197–219. https://doi.org/10.1080/19378629.2010.519773.
Galati, F., Bigliardi, B., 2019. Industry 4.0: emerging themes and future research avenues using a text mining approach. Comput. Ind. 109, 100–113. https://doi.org/10.1016/J.COMPIND.2019.04.018.
Giordano, V., Coli, E., Martini, A., 2022. An open data repository for engineering design: using text mining with open government data. Comput. Ind. 142, 103738 https://doi.org/10.1016/J.COMPIND.2022.103738.
Guo, L., Vargo, C., Pan, Z., 2016. Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. W.D.-J.& M., undefined 2016 J. Sagepub. Com. 93, 332–359. https://doi.org/10.1177/1077699016639231.
Gurcan, F., Cagiltay, N.E., 2019. Big data software engineering: analysis of knowledge domains and skill sets using LDA-based topic modeling. IEEE Access 7, 82541–82552. https://doi.org/10.1109/ACCESS.2019.2924075.
Gutierrez-Bucheli, L., Kidman, G., Reid, A., 2022. Sustainability in engineering education: A review of learning outcomes. J. Clean. Prod. 330, 129734.
Hofmann, T., 1999. Probabilistic latent semantic indexing. Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., SIGIR 1999 50–57. https://doi.org/10.1145/312624.312649.
Home - ENAEE, (n.d.). ⟨https://www.enaee.eu/⟩ (accessed April 11, 2023).
Industrial and production engineers | Esco, (n.d.). ⟨https://esco.ec.europa.eu/en/classification/occupation?uri=http://data.europa.eu/esco/isco/C2141⟩ (accessed April 12, 2023).
Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed. Tools Appl. 78, 15169–15211. https://doi.org/10.1007/S11042-018-6894-4/TABLES/11.
Johri, A., Wang, G.A., Liu, X., Madhavan, K., 2011. Utilizing topic modeling techniques to identify the emergence and growth of research topics in engineering education. Proc. - Front. Educ. Conf., FIE. https://doi.org/10.1109/FIE.2011.6142770.
Karimova, G.Z., Goby, V.P., 2020. The adaptation of anthropomorphism and archetypes for marketing artificial intelligence. J. Consum. Mark. 38, 229–238. https://doi.org/10.1108/JCM-04-2020-3785.
Kobayashi, M., Takeda, K., 2000. Information Retrieval on the Web. ACM Comput. Surv. 32, 144–173. https://doi.org/10.1145/358923.358934.
Kong, P., Cornet, H., Frenkler, F., 2018. October). Personas and emotional design for public service robots: A case study with autonomous vehicles in public transportation. In: *2018 international conference on cyberworlds (cw)*. IEEE, pp. 284–287.
D.R. Krathwohl, A Revision of Bloom's Taxonomy: An Overview, Https://Doi.Org/10.1207/S15430421tip4104_2. 41 (2010) 212–218. https://doi.org/10.1207/S15430421TIP4104_2.
Linguistic Features·spaCy Usage Documentation, (n.d.). ⟨https://spacy.io/usage/linguistic-features⟩ (accessed April 12, 2023).
(1) LinkedIn, (n.d.). ⟨https://www.linkedin.com/mynetwork/⟩ (accessed April 12, 2023).
Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W., 2016. An overview of topic modeling and its current applications in bioinformatics. Springerplus 5. https://doi.org/10.1186/S40064-016-3252-8.
Lupi, F., Mabkhot, M.M., Finżgar, M., Minetola, P., Stadnicka, D., Maffei, A., Litwin, P., Boffa, E., Ferreira, P., Podržaj, P., Chelli, R., Lohse, N., Lanzetta, M., 2022. Toward a sustainable educational engineer archetype through Industry 4.0. Comput. Ind. 134, 103543 https://doi.org/10.1016/J.COMPIND.2021.103543.
Mabkhot, M.M., Ferreira, P., Maffei, A., Podržaj, P., Mądziel, M., Antonelli, D., Lanzetta, M., Barata, J., Boffa, E., Finżgar, M., Paśko, Ł., Minetola, P., Chelli, R., Nikghadam-Hojjati, S., Wang, X.V., Priarone, P.C., Litwin, P., Stadnicka, D., Lohse, N., Lupi, F., 2021. Mapping industry 4.0 enabling technologies into united nations sustainability development goals. Vol. 13, 2560 Sustainability 2021 13, 2560. https://doi.org/10.3390/SU13052560.
Maffei, A., Boffa, E., Lupi, F., Lanzetta, M., 2022. On the design of constructively aligned educational unit. Educ. Sci. 12 (7), 438.
D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, S. Adam, Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology, Https://Doi.Org/10.1080/19312458.2018.1430754. 12 (2018) 93–118. https://doi.org/10.1080/19312458.2018.1430754.
Massey, A.K., Eisenstein, J., Antón, A.I., Swire, P.P., 2013. Automated text mining for requirements analysis of policy documents. IEEE, pp. 4–13. July.
MECHANICAL ENGINEERING | Politecnico di Torino, (n.d.). ⟨https://www.polito.it/en/education/master-s-degree-programmes/mechanical-engineering⟩ (accessed April 12, 2023).
Mechanical Engineering MEng | Undergraduate study | Loughborough University, (n.d.). ⟨https://www.lboro.ac.uk/study/undergraduate/courses/mechanical-engineering-meng/#modules_year_1⟩ (accessed April 12, 2023).
Miaskiewicz, T., Kozar, K.A., 2011. Personas and user-centered design: how can personas benefit product design processes? Des. Stud. 32, 417–430. https://doi.org/10.1016/J.DESTUD.2011.03.003.
Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 262–272. July.
models.ldamodel – Latent Dirichlet Allocation — gensim, (n.d.). ⟨https://radimrehurek.com/gensim/models/ldamodel.html⟩ (accessed April 11, 2023).
Neate Aikaterini Bourazeri Abi Roper, T., Stumpf Stephanie Wilson, S., Neate, T., Bourazeri, A., Roper, A., Stumpf, S., Wilson, S., 2019. Co-created personas: Engaging and empowering users with diverse needs within the design process. Dl. Acm. Org. 12 https://doi.org/10.1145/3290605.3300880.
Pejic-Bach, M., Bertoncel, T., Meško, M., Krstić, Ž., 2020. Text mining of industry 4.0 job advertisements. Int J. Inf. Manag. 50, 416–431. https://doi.org/10.1016/J.IJINFOMGT.2019.07.014.
Phuong, D.T.D., Harada, F., Shimakawa, H., 2013. Estimating student persona through factorization of learning portfolio. In: *2013 IEEE Region 10 Humanitarian Technology Conference*. IEEE, pp. 221–226.

Programme curriculum | Politecnico di Torino, (n.d.). ⟨https://www.polito.it/en/education/master-s-degree-programmes/engineering-and-management/programme-curriculum⟩ (accessed April 12, 2023).
Project description / Maestro, (n.d.). ⟨https://maestro.prz.edu.pl/project-description⟩ (accessed April 12, 2023).
pyLDAvis — pyLDAvis 2.1.2 documentation, (n.d.). ⟨https://pyldavis.readthedocs.io/en/latest/readme.html⟩ (accessed April 12, 2023).
Rani, S., Kumar, M., 2021. Topic modeling and its applications in materials science and engineering. Mater. Today Proc. 45, 5591–5596. https://doi.org/10.1016/J.MATPR.2021.02.313.
R. Řehůřek, P. Sojka, Software framework for topic modelling with large corpora, (2010). ⟨https://repozitar.cz/publication/15725/?lang=en;kod=S530⟩ (accessed April 12, 2023).
Sievert, C., 2014. K.S.-P. of the workshop on interactive, undefined 2014, LDAvis: a method for visualizing and interpreting topics. Aclanthology. Org. 63–70. ⟨https://aclanthology.org/W14-3110.pdf⟩ (accessed April 11, 2023).
Silva, C.C., Galster, M., Fabian Gilson, ·, De, A., Camila, L., Silva, C., Gilson, F., 2021. Topic modeling in software engineering research, 2021 26:6 Empir. Softw. Eng. 26, 1–62. https://doi.org/10.1007/S10664-021-10026-0.
Spreafico, C., Spreafico, M., 2021. Using text mining to retrieve information about circular economy. Comput. Ind. 132, 103525 https://doi.org/10.1016/J.COMPIND.2021.103525.
Università di Pisa: corso di laurea in INGEGNERIA MECCANICA, (n.d.). ⟨https://www.unipi.it/index.php/lauree/regolamento/10541⟩ (accessed April 12, 2023).
Università di Pisa: corso di laurea in INGEGNERIA GESTIONALE, (n.d.). ⟨https://www.unipi.it/index.php/lauree/regolamento/10534⟩ (accessed April 12, 2023).
Vayansky, I., Kumar, S.A.P., 2020. A review of topic modeling methods. Inf. Syst. 94, 101582 https://doi.org/10.1016/J.IS.2020.101582.
Vincent, C.J., Blandford, A., 2014. The challenges of delivering validated personas for medical equipment design. Appl. Erg. 45, 1097–1105. https://doi.org/10.1016/J.APERGO.2014.01.010.
Wood, A.E., Mattson, C.A., 2019. Quantifying the effects of various factors on the utility of design ethnography in the developing world. Res Eng. Des. 30, 317–338. https://doi.org/10.1007/S00163-018-00304-2.
World University Rankings 2021 | Times Higher Education (THE), (n.d.). ⟨https://www.timeshighereducation.com/world-university-rankings/2021/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats⟩ (accessed April 12, 2023).
Yüksel, N., Börklü, H.R., Sezer, H.K., Canyurt, O.E., 2023. Review of artificial intelligence applications in engineering design perspective. Eng. Appl. Artif. Intell. 118, 105697.
Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., Zou, W., 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinforma. 16, 1–10. https://doi.org/10.1186/1471-2105-16-S13-S8/FIGURES/6.