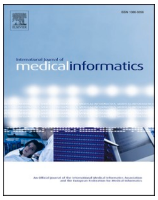
Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Medical informatics labor market analysis using web crawling, web scraping, and text mining

Jürgen Schedlbauer [a,*], Georgios Raptis [a], Bernd Ludwig [b]

[a] OTH Regensburg, Seybothstraße 2, 93053 Regensburg, Germany
[b] University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

ARTICLE INFO

Keywords:
Medical informatics
Graduate employability
Text mining
Competence-based education
Soft skills

ABSTRACT

Objectives: The European University Association (EUA) defines "employability" as a major goal of higher education. Therefore, competence-based orientation is an important aspect of education. The representation of a standardized job profile in the field of medical informatics, which is based on the most common labor market requirements, is fundamental for identifying and conveying the learning goals corresponding to these competences.

Methods: To identify the most common requirements, we extracted 544 job advertisements from the German job portal, STEPSTONE. This process was conducted via a program we developed in R with the "rvest" library, utilizing web crawling, web extraction, and text mining. After removing duplicates and filtering for jobs that required a bachelor's degree, 147 job advertisements remained, from which we extracted qualification terms. We categorized the terms into six groups: professional expertise, soft skills, teamwork, processes, learning, and problem-solving abilities.

Results: The results showed that only 45% of the terms are related to professional expertise, while 55% are related to soft skills. Studies of employee soft skills have shown similar results. The most prevalent terms were programming, experience, project, and server. Our second major finding is the importance of experience, further underlining how essential practical skills are.

Conclusions: Previous studies used surveys and narrative descriptions. This is the first study to use web crawling, web extraction, and text mining. Our research shows that soft skills and specialist knowledge carry equal weight. The insights gained from this study may be of assistance in developing curricula for medical informatics.

1. Introduction

In the Bologna process, the European University Association (EUA) defines the European Higher Education Area as an area for international cooperation in higher education. The main goals of its members are structural reforms and the exchange of tools for improving education. Easier integration into the labor market (*employability*), mobility within the EU, and lifelong learning have been established as educational goals [1]. To ensure the quality of individual courses, a continuous improvement process has been implemented; further, standards and guidelines have been defined [2]. In 2012, Rieckmann [3] identified the most relevant qualifications that students should acquire for the sustainable development of future-oriented higher education. The study confirms that competences in systemic thinking and problem solving, critical thinking, cooperation in (heterogeneous) groups, interdisciplinary work, the planning/realization of innovative projects, communication, and the use of media are important for graduates as industries, the society, and the environment advance in complexity. A detailed examination of competence-based medical education (CBME) by Holmboe et al. [4] demonstrated that CBME is an important collection of principles and approaches for teaching medical students. When implemented effectively and dynamically, CBME can improve all medical training programs, and therefore, help medical professionals serve their patients better. Although numerous studies have attempted to develop frameworks of competencies, few have addressed competence-based education in computer sciences, especially for medical informatics [5, 6]. A systematic review of studies on competence-based education in medical informatics was undertaken by Davies et al. [7] in an attempt to

* Corresponding author.
E-mail addresses: juergen.schedlbauer@extern.oth-regensburg.de (J. Schedlbauer), georgios.raptis@oth-regensburg.de (G. Raptis), bernd.ludwig@ur.de (B. Ludwig).

https://doi.org/10.1016/j.ijmedinf.2021.104453
Received 4 December 2020; Received in revised form 18 March 2021; Accepted 29 March 2021
Available online 8 April 2021

identify the core qualifications and skills of medical computer scientists. Several studies have utilized semi-structured expert interviews, manual analyses of job postings, and online surveys to create representative job profiles for medical informatics [8–10].

Due to the ever-relevant goal of employability and rapidly changing labor market, it has become increasingly important to have deep insight into current skill and knowledge requirements [11]. This study is part of a mixed method approach, comprising two questions and a hypothesis, as shown in Fig. 1 [12]. The first question: “What are the current qualification requirements (competences) of the job market?” is essential, as a graduate’s employment opportunities increase when their skills match the current requirements [13].

Second question: “Is the curriculum of medical informatics in German universities of applied sciences competence-based?”. The hypothesis is that the alignment of education with outcome-oriented competences improves employability. The present work develops a competence profile which is to serve as a basis for the analysis and improvement of a curriculum for health informatics. Due to the limited number of words in a research paper, this research will be presented in another study [14].

Job advertisements are an important data source because they contain information regarding the skills and knowledge required for specific jobs. To the best of our knowledge, this is the first study to use text mining for the analysis of job postings for medical informatics. Prior studies examined this type of data manually [15]. Text mining is becoming increasingly relevant in the analysis of job postings – more powerful hardware has enabled larger quantities of data to be sampled in smaller timeframes, enabling more job advertisements to be examined in a shorter period [16].

2. Background

2.1. Employability and soft skills

Although there is no clear definition of employability in the field of higher education, Andrews and Higson [17] list several skills and qualifications that they deem essential for graduate employability:

- Business-specific skills (Hard business-related knowledge and skills);
- Interpersonal competencies (Soft business-related skills); and
- Work experience and work-based learning.

As soft and hard skills have the same importance, universities should give both aspects similar weight when planning their education programs [18]. The main weakness of the existing International Medical Informatics Association (IMIA) competence framework for medical informatics is that it does not emphasize soft skills [19]. In terms of education, the framework focuses primarily on Biomedical and Health Informatics, secondarily on Medicine, Health System Organization, Health and Biosciences, and tertiarily, on Informatics/Computer Science, Mathematics and Biometry. Because employers emphasize the possession of soft skills as an important hiring criterion, these skills should also be taught at universities [20]. Typical workplace soft skills are teamwork, problem solving, communication, willingness to work, time management, and continuous learning.

2.2. Text mining

Recent advances in text mining methods have facilitated the investigation of job profiles, as software for web crawling, web extraction, and text analysis has become widely available. Nevertheless, these tools have only been utilized for a small number of job profiles, such as for Industry 4.0 and Business Analytics [21,22]. Since Healthcare 4.0 will lead to smarter medicine with new processes and treatment methods, it is unsurprising that workplace environments for medical informatics are changing rapidly [23]. It is therefore necessary, for the universities that offer courses in medical informatics to understand the current skill and competence requirements of the healthcare sector to continuously adapt their curriculum to the labor market.

2.3. Competence orientation and learning goals in higher education

The quality of higher education in medical informatics should be ensured through an accreditation program on an international level [24]. If the 2010 IMIA framework forms the basis for accreditation, no learning objectives containing soft skills will be included [19]. Succi and Canovi [25] suggest aligning higher education institutes with companies, enabling students to acquire and develop essential skills that will help them adapt to the labor market, thereby improving their employability. A study on competence-based education in medicine demonstrated that adding practical exercises helps to meet the needs of patients and the general public. Teaching these principles and approaches can also result in better outcomes for students [4]. Therefore, learning goals should be defined by workplace requirements.

In this study, we extracted and evaluated the skills that medical computer scientists are required to have based on job postings from an online job portal, and then, categorized the extracted key terms as hard or soft skills to model a standardized job profile.

3. Material and methods

The quality and quantity of web data available for automated retrieval are both growing continuously, making web data a valuable source of information [26]. We extracted the data used in our study via script we developed (in the programming language R), utilizing the rvest and tidytext packages [27,28]. To understand whether the expectations of employers are reflected in job postings, we used a combination of the established methods of content analysis and text mining for documents and web pages, respectively. Content analysis is typically performed in three stages: preparation, organization, and reporting [29].

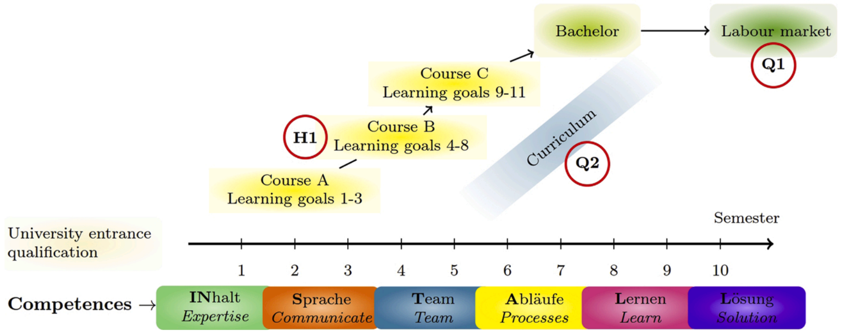


Fig. 1. Optimization of medical informatics higher education curriculum to improve employability.

3.1. Preparation

The preparation phase handles the acquisition of representative data that can help answer the research question. We believe that job postings from the leading German online job portal, STEPSTONE, are a good data source. By utilizing web crawling, we collected relevant job advertisements, posted between February and March 2020, [30,31]. The selection criteria were the German search terms *Medizin* and *Informatik*. The collection process was performed three times (with 173, 175, and 196 postings per iteration), resulting in 544 total job postings. Fig. 2 illustrates our data refinement process.

First, we eliminated 300 duplicates within the extracted timeframe using the job IDs in the advertisements. We then removed 97 entries for jobs that did not require a bachelor's or masters's degree, leaving a final group of 147 job posting documents with an HTML document structure comprising six data fields:

- Unique job ID
- Company name
- Title
- Your profile (the applicant's qualifications)
- Tasks (job description)
- Conditions

To validate the data, we performed a frequency analysis of company names using a word cloud consisting of the top 50 companies searching for employees [32]. The company names are visualized in Fig. 3, which covers 97 of the 147 (66%) evaluated job postings.

We used an inductive content analysis approach to ensure that the samples only comprised job advertisements from the healthcare sector. We identified relevant employers using the company description to create the following business area categories [33]:

- Hospital
- Foundation/institute
- Medical technology manufacturer
- Physician practice
- Recruiter
- Nursing home
- Pharmacy/pharma
- Software manufacturer
- Laboratory
- Academy
- Biotechnology
- Consulting
- Health management
- Insurance company
- Not applicable

The top 50 companies covering 97 of the 147 (66%) job advertisements in the frequency analysis of company names are visualized in Fig. 3 as Wordcloud [32].

3.2. Organization phase

In the organizational phase of our deductive approach, we developed a competence matrix based on research that examined the requirements of the medical informatics labor market. A considerable volume of literature has been published on skills in relation to employability [34,25,17,20]. These studies show that companies expect graduates to have a mix of hard and soft skills. The most valuable skills we identified for medical informatics are

- Expertise
- Communication
- Teamwork
- Processes
- Lifelong learning
- Problem solving

The script we developed for this study uses text mining to extract key words from the *Your Profile* section of job postings, and then, classifies them into the six categories that we discovered to be most relevant to employability. Key word extraction is essential for this process, as it includes a text pre-processing operation, typically comprising tokenization, filtering, lemmatization, and stemming tasks [26].

Tokenization refers to breaking a word sequence into smaller pieces (words), and concurrently, removing special characters and punctuation. After breaking up a sequence, we used the anti_join function in R to filter for stop words. This function utilizes an existing German stop word

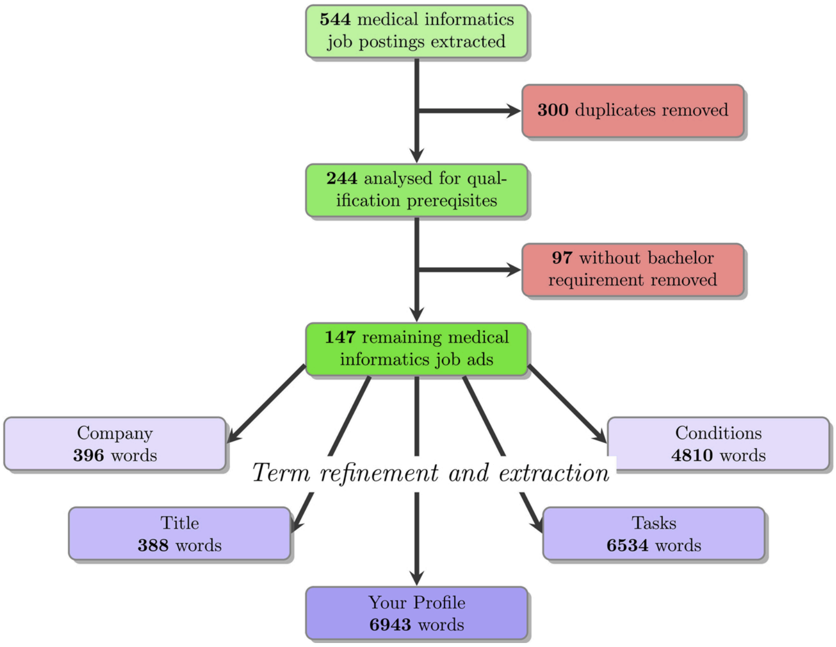

544 medical informatics job postings extracted
300 duplicates removed
244 analysed for qualification prereqisites
97 without bachelor requirement removed
147 remaining medical informatics job ads
Term refinement and extraction
Company **396** words
Conditions **4810** words
Title **388** words
Tasks **6534** words
Your Profile **6943** words

Fig. 2. Extraction and refinement of job postings.

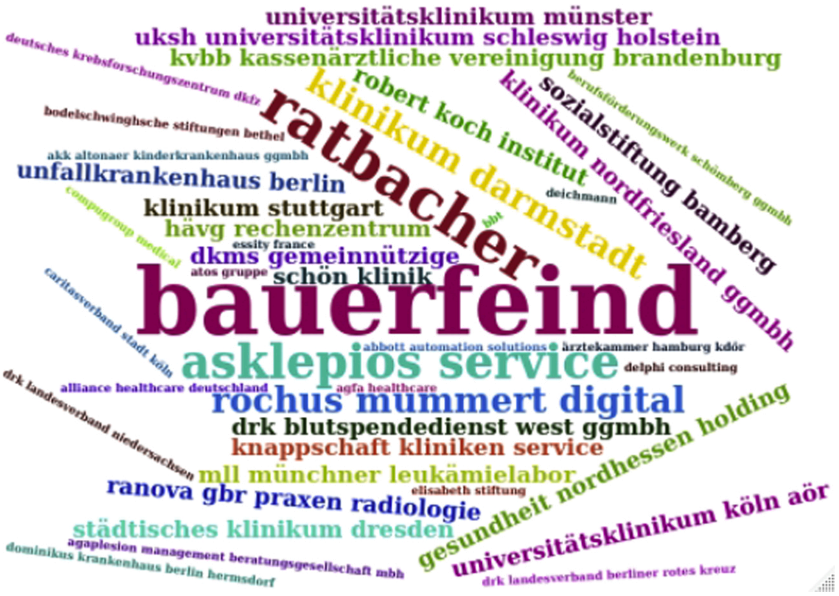


Fig. 3. Top 50 Word cloud: Companies searching for medical informatics employees.

list, which we extended with our own list of words, not covered by the base list. The examples of terms we added are *interest*, *bring along*, *comparable*, *the same*, and *desirable* – despite being typical in job advertisements, we did not consider them necessary for the skill category mapping process.

Lemmatization and stemming are used to group synonyms and words with similar subcomponents into a single item. As German medical informatics job advertisements often contain a mixture of German and English wording, we encountered two major problems. First, there is no German lemmatization catalog. Second, the algorithm was unable to function correctly in the presence of two distinct languages. Therefore, to identify similar words, we created our own bilingual catalog based on the optimal string-alignment-distance function, a straightforward extension of the Levenshtein distance that allows for the transposition of adjacent characters [35]. An example of this is *Teamfähigkeit*, which can be identified with *team(10)*, *teamgeist(7)*, *teams(10)*, *teamorientiertes (11)*, *teamorientierten(11)*, *teamorientiert(9)*, *teamorientierte(10)*, *teamarbeit(7)*, *teamspieler(8)*.

After the final processing step, 1533 single words and 4646 two-word phrases remained in the *Your Profile* section. To determine the weighting of the words, the term frequency-inverse document frequency (TF-IDF) was calculated with the *bind_tf_idf* function [36], and the most “important” 100 terms are divided among the six competence areas. The TF-IDF is calculated by multiplying the term frequency (TF) by the inverse document frequency (IDF) (TF-IDF = TF * IDF). The TF measures the frequency of a term within a document. For example, in a job advertisement (100 words), four uses of the term *experience* results in a TF value of 0.04 for *experience*. Because documents are rarely of the same length, longer texts can skew this value as they invariably include more occurrences of common terms. The IDF offsets this skew by assigning a lower weight to frequently occurring words and a greater weight to infrequently occurring ones using the logarithm function: $\log_e$(total words/ number of instances of a given word) [37].

3.3. *Reporting phase*

Descriptive statistics were generated for the categorized companies, profile terms, and title to ensure that the analyzed job advertisements belonged to the healthcare and medical informatics sectors. We analyzed the validity of the six categories of hard and soft skills using a treemap graph of the TD_IDF in job advertisements that belonged to a given category [38]. A treemap graph is a special method for visualizing large hierarchical data sets.

As shown in Table 1, over half of the employers can be classified as hospital (50.34%). When grouped together, foundation/institute, physician practices and medical technology manufacturer constitute 25,84% of the employers, and seem to carry greater weight than the remaining 23,82% of categories, which include a further 11 business areas.

Descriptive statistics for terms in job titles are displayed in Table 2. The most interesting aspects of these results are the high frequency of occurrences of the term *administration of IT systems* and the frequent mention of the software manufacturer SAP in 11 of 147 job advertisements. Developer, project manager, and manager are also among the most important positions. These results show that the focus of a job title tends to be the task of the job, not the skills required to accomplish it.

4. Results

First, we present an analysis of a representative example of a job posting, followed by a descriptive analysis of the *Your Profile*, *Title*, and *Company* sections. Finally, we present a map and visualization of the key terms for the six hard and soft skill categories using the TF-IDF function and show the top 100 terms and their categories in a bigram graph.

Table 1
Categorization of 147 health care companies by sector and the corresponding number of job advertisements.

Sector	*n*	%
Hospital	74	50.34
Foundation/institute	15	10.20
Medical technology manufacturer	12	8.16
Physician practice	11	7.48
Recruiter	9	6.12
Nursing home	5	3.40
Pharmacy/pharma	4	2.72
Software manufacturer	4	2.72
Laboratory	3	2.04
Academy	2	1.36
Biotechnology	2	1.36
Consulting	2	1.36
Health management	2	1.36
Insurance company	1	0.68
Not applicable	1	0.68

Table 2
Top 10 terms acquired via frequency analysis (388 terms) of job titles.

Term	*n*	%
systemadministrator	20	5.38
sap	11	2.96
administrator	9	2.42
projektmanager	9	2.42
informatiker	8	2.15
entwickler	7	1.88
system	7	1.88
manager	6	1.61
netzwerkadministrator	6	1.61
softwareentwickler	6	1.61

4.1. *Representative job posting*

A typical example of a job advertisement (jobid = 711) can be seen in Table 3. The 10 most commonly mentioned terms (n) are listed with their calculated TF- and TF_IDF values. In total, 85 terms consisting of 67 unique words were used in this job advertisement. The highest TF value were assigned to Berufserfahrung (work experience), which was mentioned 10 times, and Betrieb (operation), with four occurrences. Terms related to healthcare, namely, Krankenhaus (hospital) and Medizintechnik (medical technology), were used twice. When calculated over all 147 postings, the TF_IDF exhibits a different focus: the top terms are Betrieb (operations, 0.09), Medizintechnik (medical technology, 0.08), and komplexer (more complex, 0.05).

4.2. *Descriptive statistics*

4.2.1. *Company names*

Table 1 shows that half of the job advertisements were published by hospitals, while other healthcare sectors such as foundations/institutes (10.2%), medical technology manufacturers (8.16%), and physician practices (8.16%), together represent only a quarter of the 147 total advertisements. Surprisingly, a few advertisements came from pharmacies (2.7%), software manufacturers (2.72%), and laboratories (2.04%). Fig. 4 lists German healthcare expenses for hospitals (25%), physical practices (14%), pharmacies (13%), and nursing homes (15%). The comparison of the percentages of sectors from Table 1 and Fig. 4 shows that their percentages are similar, as an exception; however, large deviations are observed particularly between those for hospitals (double) and pharmacies (20%) [39].

In the job titles, 9% of the terms refer to activities related to system administration (systemadministrator, administrator, netzwerkadministrator, system) (see Table 2). Software development is mentioned as entwickler (1.88%) and softwareentwickler (1.61%). Management positions are also commonly mentioned, with companies searching for employees in management (1.61%) or project management (2.42%). Of particular interest is the frequent mention of the software manufacturer SAP, comprising three percent of the 388 terms used in the titles.

Table 3
Representative job advertisement: Top 10 words mentioned in *Your Profile* (including TF, IDF, and TF-IDF).

jobid	Terms	*n*	tf	idf	tf_idf
711	berufserfahrung	10	0.12	0.09	0.01
711	betrieb	4	0.05	1.81	0.09
711	fundierte	2	0.02	1.52	0.04
711	kenntnisse	2	0.02	0.34	0.01
711	komplexer	2	0.02	2.05	0.05
711	krankenhaus	2	0.02	1.90	0.04
711	medizintechnik	2	0.02	3.38	0.08
711	wünschenswert	2	0.02	1.44	0.03
711	abläufe	1	0.01	1.44	0.02
711	arbeitsweise	1	0.01	0.82	0.01

4.2.2. *Your profile*

On average, 47.2 unique words were used in the *Your profile* section. Notably, in Table 4, the term Berufserfahrung (work experience) is used 298 times, an average of two times per job advertisement. Some general words describing the applicant's knowledge/skill level are kenntnisse (knowledge, 192) and gute (good, 127). The bigram analysis showed that the terms *gute kenntnisse* (good knowledge, 49) and *fundierte kenntnisse* (sound knowledge, 24) are used 73 times (see Table 5). As expected, a degree (abgeschlossenes 148, studium 128) in computer science (107) is a frequently requested qualification. Companies were primarily interested in soft skills such as teamfähigkeit (teamwork, 89) and kommunikationsfähigkeit (communication skills, 77). A decent grasp of a language is also addressed in the bigram table, as *wort schrift* (word writing) and *gute deutschkenntnisse* (good knowledge of German) are among the top 10 entries.

4.3. *Analysis of soft and hard skills*

To compare the expectations of employers for hard and soft skills, we grouped the terms in *Your profile* into six categories: expertise, communication, teamwork, processes, lifelong learning, and problem solving. The additional category *not applicable* (NA) serves as an umbrella term for all other miscellaneous terms. The two terms with the highest TF_IDF in each category are shown in Table 6.

The value for experience (3.33) is notable, which is in the learning category. The remaining top 100 terms have TF_IDF ranging between 0.46 (commitment) in the team category and 1.89 (programming) in the expertise category.

The treemap graph in Fig. 5 indicates that expertise 43.3% (summary of TF_IDF = 39.7) has the highest proportion among the six skill groups, followed by learning (20%), solution (11.7%), team (9.21%), communication (9.20%), and processes (6.6%) [38].

The core terms in the expertise category are focused on software development, with expressions such as programming (TF_IDF 1.89), java (0.86), abap (0.51), html (0.48), and applications (0.66) being used most frequently. Several other notable subcategories of expertise are

- Administration (1.30) of servers (1.61), network technology (1.34), clients (0.66), LAN (0.54), VMWare (0.70), virtualization (0.58), and general systems (0.94);
- Databases (1.06) with sql (1.12), oracle (0.80), and data (0.57) in general; and
- Software such as SAP (1.58) and Microsoft (1.24), including office (0.93), active directory (0.59), exchange (0.49), and sharepoint (0.42).

Surprisingly, hospital information systems (0.5) are mentioned infrequently, with only the leading German market provider *agfa orbis* being included in the top 100 terms.

Employers view acquired learning as a secondary requirement, and instead emphasize that experience (3.35) and several years (1.35) of work experience (1.16) are important for employment. Further qualifications primarily describe a prospective employee's attitude toward work, employing terms such as self-reliant (1.2), goal-oriented (0.63), engagement (0.76), joy (0.61), and interest (0.61). These characteristics also reflect the willingness to learn at work. Terms, such as think (0.92) and comprehension (0.58), hint at an expectation to not only follow orders, but also be innovative.

Teamwork, along with the accompanying social soft skills, is a standard requirement in any workplace. Dealing (1.29) with other people, cooperation (0.48), and social (0.47) are terms commonly used to describe the interaction between colleagues. To be customer oriented (kundenorientiert, 0.74), service oriented (serviceorientiert, 0.56), and service orientation (dienstleistungsorientiert, 0.60) are the most common terms related to customer satisfaction in the service sector.

Implementing solutions is usually performed in projects (1.72), and

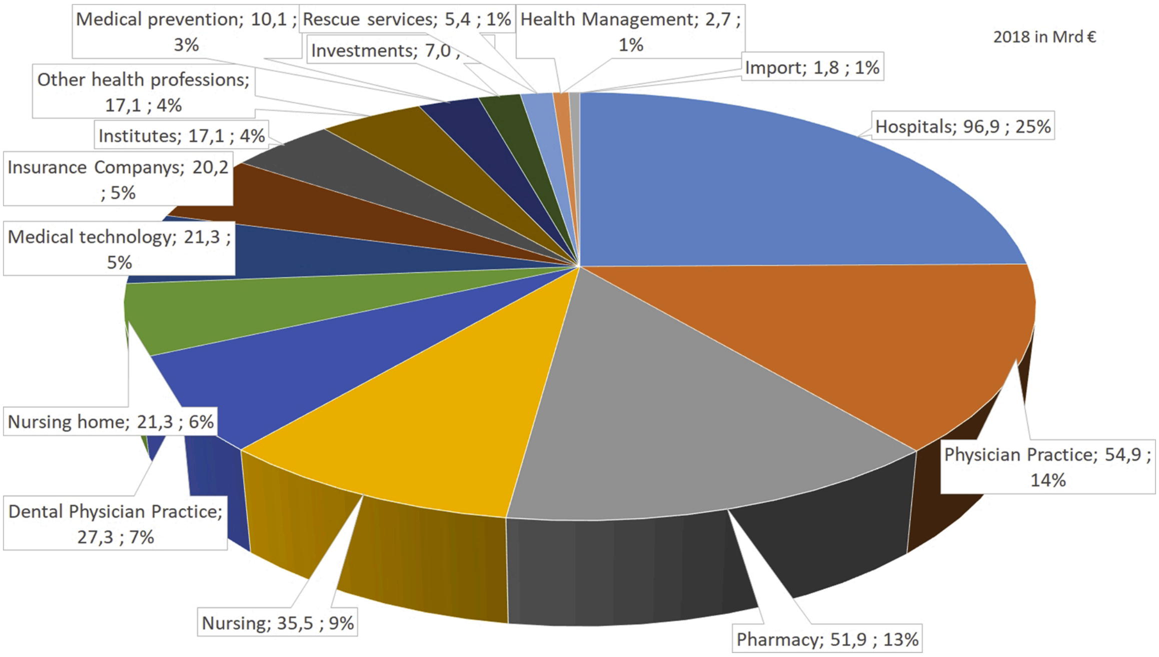


Fig. 4. Healthcare spending in Germany by sector in 2018 (€390 billion).

Table 4
Top 10 terms extracted from *Your profile* (6943 terms).

Term	*n*	%
berufserfahrung	298	4.29
kenntnisse	192	2.77
abgeschlossene	148	2.13
studium	129	1.86
gute	127	1.83
informatik	107	1.54
teamfähigkeit	89	1.28
profil	78	1.12
kommunikationsfähigkeit	77	1.11
ausbildung	74	1.07

Table 5
Top 10 bigrams extracted from the qualification profile of job advertisements.

Bigram	*n*	%
abgeschlossene studium	84	1.24
gute kenntnisse	49	0.72
studium informatik	49	0.72
mehrjährig berufserfahrung	41	0.60
profil abgeschlossene	40	0.59
erfolgreich abgeschlossene	36	0.53
wort schrift	27	0.40
vergleichbare ausbildung	26	0.38
fundierte kenntnisse	24	0.35
gute deutschkenntnisse	24	0.35

Table 6
Top two terms for each skill category (including NA) in terms of the German–English category TF_IDF.

German	English	Category	TF_IDF
gute	Quality	NA	1.75
profil	Profile	NA	1.45
programmierung	Programming	Expertise	1.89
netzwerktechnik	Network	Expertise	1.34
erfahrung	Experience	Learn	3.33
mehrjährig	Perennial	Learn	1.35
projekt	Project	Solution	1.72
lösungsorientiert	Solution-oriented	Solution	1.18
abläufe	Processes	Processes	1.29
strukturiert	Structured	Processes	1.20
englisch	English	Communication	1.19
gesundheitswesen	Healthcare	Communication	1.17
umgang	Handling	Team	1.29
teamfähigkeit	Teamwork	Team	1.08

requires solution orientation (1.18) as well as analytical (1.07) and conceptual (1.05) thinking. On a higher level, leading several projects requires the management (0.71) of resources, thinking strategically (0.64), and planning (0.57).

Communication skills, for example, a good grasp of languages such as English (1.18) and German (1.10), are essential for understanding customers. The words, medicine (1.05), hospital (0.97), and clinic (0.84), express the desire for employees to grasp the technical vocabulary used in health care.

Furthermore, process (1.28) knowledge is necessary for medical informatics because software used in this field commonly implements working processes (0.71). These are embedded in structured (1.19) and systematically (0.54) formed analyses of work areas.

Closer inspection of the NA category in Table 7 reveals that the terms in this category either refer to the possession of knowledge (qualification, have, comparable, bring along, profile) or the level of skills (good, high).

5. Discussion

The objective our study is to optimize the medical informatics education at universities of applied sciences in terms of the requirements of the job market. The first step involved ascertaining whether the hard and soft skills mentioned in employer surveys are represented in job postings for medical informatics since soft skills are not emphasized in the recommendations of professional associations for medical informatics [6,7].

5.1. Principal findings

Our primary findings were: (1) web crawling combined with text mining and content analysis is effective in identifying job profiles; (2)

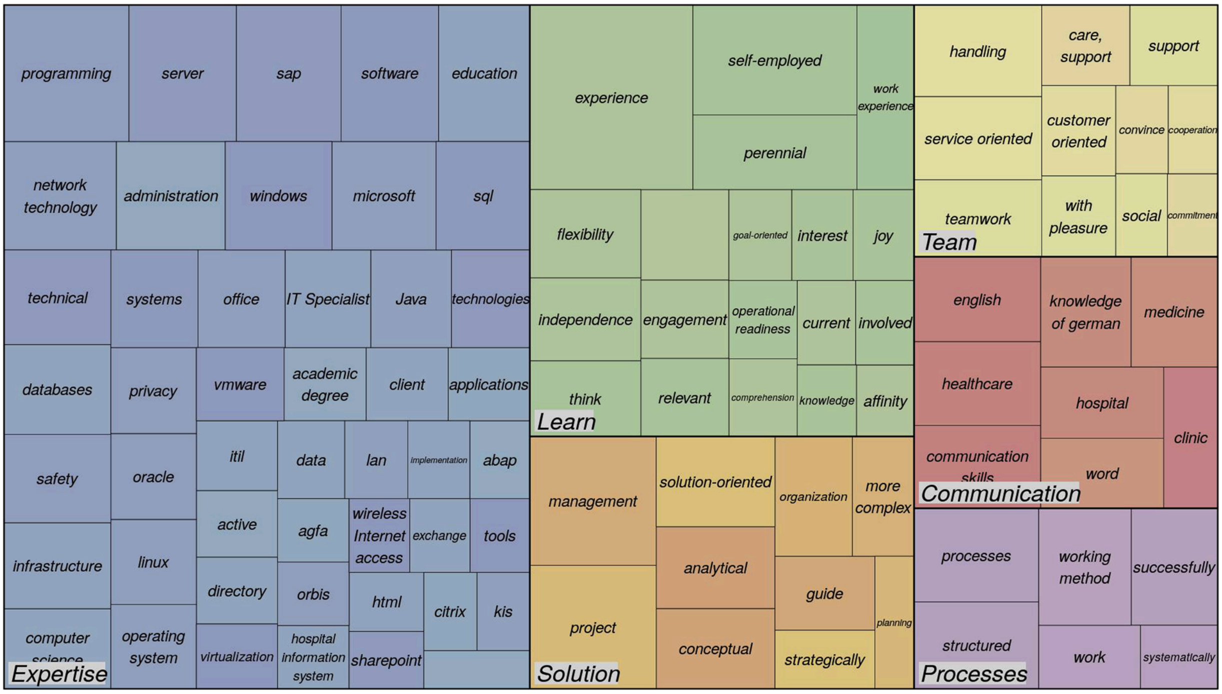


Fig. 5. Treemap representation of the six competence areas and top 100 TF_IDF terms in job advertisements (translated to English).

Table 7
Top 10 terms in the *Not Applicable* category of the *Your Profile* section.

German	English	Category	TF_IDF
gute	Good	NA	1.75
profil	Profile	NA	1.45
verfügen	Have	NA	1.35
idealerweise	Ideally	NA	1.34
kenntnisse	Knowledge	NA	1.30
vergleichbare	Comparable	NA	1.22
bringen	Bring along	NA	1.21
hohe	High	NA	1.18
umfeld	Environment	NA	1.16
qualifikation	Qualification	NA	1.15

hospitals are the main employers searching for employees with a degree in medical informatics; (3) terms such as work experience and good or sound knowledge were used in every job advertisement, indicating that almost no entry level positions are advertised; (4) in contrast to earlier findings, soft skills comprise more than half of the requirements; and (5) the administration and development of software are the main hard skills required for medical informatics.

5.1.1. *Semi-automated evaluation of job profiles*

In accordance with our results, previous studies have demonstrated that text mining is an effective method for analyzing job profiles since considerable information from web pages can be collected automatically [21,22]. Automatic collection is necessary because Healthcare 4.0 has already led to several changes in the workplace and will continue to do so, particularly in relation to the required skillsets [23]. Curriculum therefore needs to be adjusted at regular intervals. Nevertheless, the aggregation of terms into categories must be interpreted with caution because it is subjective and dynamic.

5.1.2. *Skill areas*

The main question we intended to answer with this study is what skills should be taught in medical informatics education to enhance employability. We found that the main hard skill requirements are administration (of networks, servers, and applications), software development, database knowledge, and the support of industry-standard applications (SAP and Microsoft). Soft skills such as learning, teamwork, communication, problem solving, and processes are of similar importance. These results echo those of Balcar [18] and Chhinzer and Russo [20], who also found that soft skills carry the same weight as hard skills for employability. The investigation conducted in this study differs from those of previous studies in that it allots soft skills to not only one or two but six different categories [17]. We believe that these six areas cover the most important aspects of employer requirements when defining the learning goals of the curriculum.

5.1.3. *Work experience*

By summing the TD_IDF of the term *experience* with its variations *work experience* and *several years*, we found a weight of 5.84, which by far exceeds the weights of other terms. This finding is consistent with that of Andrews and Higson [17], which reported that employers and graduates emphasize real work experience, especially work-based learning programs and formal placement, making them critical for finding a job. We thus suggest that in addition to the theoretical knowledge imparted at a university of applied sciences, real work processes should be practiced.

5.2. *Limitations*

The first limitation of the study is the short timeframe within which data was collected, together with the fact that we acquired it from only one source/job portal. This limitation can easily be overcome in future studies by analyzing job advertisements over a longer period and querying several job portals. The second limitation arises from the methodology of our approach. The lemmatization of German words is not an established method, because of which we used custom-developed word mapping. Moreover, we used only the top 100 terms to categorize soft and hard skills. There are likely synonyms that would alter the number and naming of categories. A further weakness is that the content analysis coding was based on the interpretation of a single researcher, leading to a potential bias. We attempted to overcome this limitation by reviewing and validating the categories several times. Although the Web Crwaling and Text Minig methodology is only used for German job advertisements, it can be used for all countries / languages. It is well known that the skills of healthcare workers in health informatics vary widely across the EU, the US, Australia, the Middle East and Africa. Therefore, further studies are needed to add knowledge about competence profiles in these countries that could be used to decide whether the IMIA recommendations should be adjusted [40].

6. Conclusion

The aim of our research was to examine the terms related to employability in medical informatics as a goal of the Bologna Process to align learning objectives with the needs of the labor market. As Healthcare 4.0 refashions what knowledge and skills are relevant, regular reviews and adaptations of curriculum are necessary. Our study presents a semi-automatic procedure of web crawling job advertisements, extracting data structures with text mining, and categorizing skills using content analyses, eventually clearly illustrating the relevant job skills within a treemap graph. The study contributes to the existing knowledge in "Recommendations on education in medical informatics":

(i) by proposing to expand the existing framework conditions and job profiles to include soft skills;
(ii) by providing current information to higher educational institutions on how to update their curricula;
(iii) by demonstrating the usefulness of the text mining approach for the exploratory analysis of job advertisements;
(iv) by revealing that hospitals are the main target group of companies hiring medical IT specialists; and
(v) by suggesting that competence-based education be supplemented with real practical exercises that reflect work in the departments of health facilities.

Further research should be carried out to examine how these recommendations are consistent with the IMIA knowledge base and with the development of medical informatics curricula in other countries [41,40,14].

Summary table

What was already known:

- Job profiles for medical informatics were created with data from expert interviews, manual content analysis of job postings, and structured questionnaires.
- Existing frameworks for medical informatics education focus on hard skills.

What this study adds:

- The combination of web crawling, text mining, and content analysis to define a medical informatics job profile.
- Soft skills such as learning, teamwork, communication, problem solving, and processes are shown to be more important than previously known.

Authors' contribution

Jürgen Schedlbauer: conceptualization, methodology, software, writing – original draft. Georgios Raptis: resources, methodology, supervision, writing – review & editing. Bernd Ludwig: methodology, supervision, funding acquisition, writing – review & editing.

Funding

This work was supported by the Bavarian Academic Forum (BayWISS) – Doctoral Consortium "Health Research," and is funded by the Bavarian State Ministry of Science and the Arts.

Conflict of 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.

Appendix A

Supplementary data associated with this article can be found in the online version at Mendeley [42].

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijmedinf.2021.104453.

References

[1] D. Crosier, A. Horvath, V. Kerpanova, D. Kocanova, T. Parveva, S. Dalferth, et al., The European Higher Education Area in 2012: Bologna Process Implementation Report, ERIC, 2012. ISBN 9292012568, https://eric.ed.gov/?redir=http.

[2] ENQA, Standards and Guidelines for Quality Assurance in the European Higher Education Area, 2015. http://www.enqa.eu/index.php/home/esg/.

[3] M. Rieckmann, Future-oriented higher education: which key competencies should be fostered through university teaching and learning? Futures 44 (2) (2012) 127–135, https://doi.org/10.1016/j.futures.2011.09.005.

[4] E.S. Holmboe, J. Sherbino, R. Englander, L. Snell, J.R. Frank, A call to action: the controversy of and rationale for competency-based medical education, Med. Teacher 39 (6) (2017) 574–581, https://doi.org/10.1080/0142159X.2017.1315067.

[5] N. Sutcliffe, S.S. Chan, M. Nakayama, A competency based msis curriculum, J. Inform. Syst. Educ. 16 (3) (2020) 8.

[6] Q.R. Huang, Competencies for graduate curricula in health, medical and biomedical informatics: a framework, Health Informatics J. 13 (2) (2007) 89–103, https://doi.org/10.1177/1460458207076465.

[7] A. Davies, J. Mueller, G. Moulton, Core competencies for clinical informaticians: a systematic review, Int. J. Med. Informatics 141 (2020) 104237, https://doi.org/10.1016/j.ijmedinf.2020.104237.

[8] M.A. Meyer, Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings, J. Am. Med. Informatics Assoc. 26 (5) (2019) 383–391, https://doi.org/10.1093/jamia/ocy181.

[9] J. Thye, T. Shaw, J. Hüsers, M. Esdar, M. Ball, B. Babitsch, et al., What are inter-professional ehealth competencies? Stud. Health Technol. Informatics 253 (2018) 201–205.

[10] M. Riley, K. Robinson, N. Prasad, B. Gleeson, E. Barker, D. Wollersheim, et al., Workforce survey of australian graduate health information managers: employability, employment, and knowledge and skills used in the workplace, Health Inform. Manag. J. 49 (2–3) (2019) 88–98, https://doi.org/10.1177/1833358319839296.

[11] N. Ashrafi, J.P. Kuilboer, C. Joshi, I. Ran, P. Pande, Health informatics in the classroom: an empirical study to investigate higher education's response to healthcare transformation, J. Inform. Syst. Educ. 25 (4) (2014) 5.

[12] R.K. Yin, Case Study Research and Applications: Design and Methods, 6th ed., SAGE, Los Angeles and London and New Delhi and Singapore and Washington DC and Melbourne, 2018. ISBN 9781506336169.

[13] Y. Kino, H. Kuroki, T. Machida, N. Furuya, K. Takano, Text analysis for job matching quality improvement, Proc. Comput. Sci. 112 (2017) 1523–1530, https://doi.org/10.1016/j.procs.2017.08.054.

[14] G. Wright, F. Verbeke, M. Nyssen, H. Betts, Health informatics: developing a master's programme in Rwanda based on the imia educational recommendations and the imia knowledge base, Stud. Health Technol. Informatics (2015) 216, https://doi.org/10.3233/978-1-61499-564-7-525.

[15] J.M. Raymond, B. Razvan, Mining knowledge from text using information extraction, SIGKDD Explor. Newsl. 7 (1) (2005) 3–10, https://doi.org/10.1145/1089815.1089817.

[16] G.K. Palshikar, R. Srivastava, S. Pawar, S. Hingmire, A. Jain, S. Chourasia, et al., Analytics-led talent acquisition for improving efficiency and effectiveness, in: A. K. Laha (Ed.), Advances in Analytics and Applications, Springer Singapore, Singapore, 2019, pp. 141–160, https://doi.org/10.1007/978-981-13-1208-3_13. ISBN 978-981-13-1208-3.

[17] J. Andrews, H. Higson, Graduate employability, 'soft skills' versus 'hard' business knowledge: a European study, Higher Educ. Europe 33 (4) (2008) 411–422, https://doi.org/10.1080/03797720802522627.

[18] J. Balcar, Is it better to invest in hard or soft skills? Econ. Labour Relat. Rev. 27 (4) (2016) 453–470, https://doi.org/10.1177/1035304616674613.

[19] J. Mantas, E. Ammenwerth, G. Demiris, A. Hasman, R. Haux, W. Hersh, et al., Recommendations of the international medical informatics association (imia) on

education in biomedical and health informatics: First revision, Methods Inform. Med. 49 (2010) 105–120, https://doi.org/10.3414/ME5119.

[20] N. Chhinzer, A.M. Russo, An exploration of employer perceptions of graduate student employability, Educ. Train. 60 (1) (2018) 104–120, https://doi.org/10.1108/ET-06-2016-0111.

[21] M. Pejic-Bach, T. Bertoncel, M. Meško, Ž. Krstić, Text mining of industry 4.0 job advertisements, Int. J. Inform. Manag. 50 (2020) 416–431, https://doi.org/10.1016/j.ijinfomgt.2019.07.014.

[22] Z.H. Przasnyski, K.C. Seal, L.A. Leon, I. Wiedenman, Skills and competencies required for jobs in business analytics: a content analysis of job advertisements using text mining, Int. J. Business Intell. Res. 8 (1) (2017) 1–25, https://doi.org/10.4018/IJBIR.2017010101.

[23] C. Chen, E.W. Loh, K.N. Kuo, K.W. Tam, The times they are a-Changin' – healthcare 4. 0 is coming!, J. Med. Syst. 44 (2) (2019) 40, https://doi.org/10.1007/s10916-019-1513-0.

[24] A. Hasman, J. Mantas, Imia accreditation of health informatics programs, Healthc. Informatics Res. 19 (3) (2013) 154–161, https://doi.org/10.4258/hir.2013.19.3.154.

[25] C. Succi, M. Canovi, Soft skills to enhance graduate employability: comparing students and employers' perceptions, Stud. Higher Educ. 45 (9) (2020) 1834–1847, https://doi.org/10.1080/03075079.2019.1585420.

[26] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. Trippe, J. Gutierrez, et al., A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques, 2017. arXiv:1707.02919.

[27] H. Wickham, M.H. Wickham, Package 'rvest'. Online-Document, 2016. https://cran.r-project.org/web/packages/rvest/rvest.pdf.

[28] Julia Silge, David Robinson, tidytext: text mining and analysis using tidy data principles in r, J. Open Source Softw. 1 (3) (2016) 37, https://doi.org/10.21105/joss.00037.

[29] S. Elo, H. Kyngäs, The qualitative content analysis process, J. Adv. Nurs. 62 (1) (2008) 107–115, https://doi.org/10.1111/j.1365-2648.2007.04569.x.

[30] L.M. Kureková, M. Beblavý, A. Thum-Thysen, Using online vacancies and web surveys to analyse the labour market: a methodological inquiry, IZA J. Labor Econ. 4 (1) (2015), https://doi.org/10.1186/s40172-015-0034-4.

[31] J. Faberman, M. Kudlyak, What does online job search tell us about the labor market? Econ. Perspect. 40 (1) (2016) 1–15.

[32] F. Heimerl, S. Lohmann, S. Lange, T. Ertl, Word cloud explorer: text analytics based on word clouds, 2014 47th Hawaii International Conference on System Sciences (2014) 1833–1842, https://doi.org/10.1109/HICSS.2014.231. ISBN 978-1-4799-2504-9.

[33] U.H. Graneheim, B.M. Lindgren, B. Lundman, Methodological challenges in qualitative content analysis: a discussion paper, Nurse Educ. Today 56 (2017) 29–34, https://doi.org/10.1016/j.nedt.2017.06.002.

[34] S. Krone, C. Patscha, M. Ratermann-Busse, F. Turber, Zukunftsperspektiven im tertiären bereich der beruflichen bildung, 2019, p. 2040. http://www.iaq.uni-due.de/iaq-forschung/2019/fo2019-02.pdf.

[35] M.P.J. van der Loo, The stringdist package for approximate string matching, R J. 6 (1) (2014) 111–122.

[36] H.C. Wu, R.W.P. Luk, K.F. Wong, K.L. Kwok, Interpreting tf-idf term weights as making relevance decisions, ACM Trans. Inform. Syst. 26 (3) (2008) 1–37, https://doi.org/10.1145/1361684.1361686.

[37] S. Qaiser, R. Ali, Text mining: Use of tf-idf to examine the relevance of words to documents, Int. J. Comput. Appl. (2018) 181, https://doi.org/10.5120/ijca2018917395.

[38] Benjamin B. Bederson, Ben Shneiderman, Martin Wattenberg, Ordered and quantum treemaps: making effective use of 2d space to display hierarchies, ACM Trans. Graph. 21 (4) (2002) 833–854, https://doi.org/10.1145/571647.571649.

[39] Destatis, Gesundheitsausgaben, 2019. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Gesundheitsausgaben/_inhalt.html.

[40] G. Wright, The development of the imia knowledge base: original research, South Afr. J. Inform. Manag. 13 (1) (2011) 1–5, https://doi.org/10.10520/EJC46343.

[41] P.L. Elkin, S.H. Brown, G. Wright, Biomedical informatics: we are what we publish, Methods Inform. Med. 52 (06) (2013) 538–546.

[42] J. Schedlbauer, G. Raptis, B. Ludwig, Medical Informatics Labour Market Analysis Using Web Crawling, -Scraping and Text Mining: Dataset v1, 2020, https://doi.org/10.17632/rjkvnhpzyz.1.