Clean Technologies and Environmental Policy (2021) 23:1215–1231
https://doi.org/10.1007/s10098-020-02000-z

ORIGINAL PAPER

# Building information modelling knowledge harvesting for energy efficiency in the Construction industry

**Andrei Hodorog[1] · Ioan Petri[1] · Yacine Rezgui[1] · Jean-Laurent Hippolyte[1]**

Received: 12 August 2020 / Accepted: 18 November 2020 / Published online: 6 December 2020


## Abstract

The recent adoption of building information modelling (BIM), and the quest to decarbonise our built environment, has impacted several segments of the supply chain, including design and engineering practitioners, prompting the need to redefine the construction personnel positions along with associated skills and competencies. The research informs ways in which practitioners can fully embrace the potential of BIM for energy efficiency to promote sustainable interventions by improving existing training practices and identifying new training requirements as BIM evolves and as practitioners' ICT (Information and Communications Technology) maturity levels improve. This is achieved by adopting a novel text-mining approach which analyses social media alongside secondary sources of evidence to establish a level of correlation between BIM roles and skills. The use of ontological dependency analysis has helped to understand the degree of correlation of skills with roles as a method to inform training and educational programmes. A key outcome from the research is a semantic web-based mining environment which determines BIM roles and skills, as well as their correlation factor, with an application for energy efficiency. The paper also evidences that (a) construction skills and roles are dynamic in nature and evolve over time, reflecting the digital transformation of the Construction industry, and (b) the importance of socio-organisational aspects in construction skills and related training provision.

✉ Ioan Petri
petrii@cardiff.ac.uk

Andrei Hodorog
hodoroga@cardiff.ac.uk

Yacine Rezgui
rezguiy@cardiff.ac.uk

Jean-Laurent Hippolyte
hyppolytej@cardiff.ac.uk

[1] BRE Institute of Sustainable Engineering, School of Engineering, Cardiff University, 52 The Parade, Cardiff, UK

**Graphic abstract**


**Keywords** Data mining · Construction digitalisation · Building information modelling · Energy efficiency · Skills · Roles

## Introduction

Construction is an information-intensive industry. This wealth of information is used to design, construct, operate, and decommission buildings. The advent of the Internet and its by-products, mainly social media and the IoT (Internet of Things), have dramatically expanded this volume of data and information. This presents a unique opportunity for data analysis and interpretation to improve the quality, sustainability, and resilience of our buildings. Text-mining and clustering techniques provide the opportunity to enhance the understanding of the implications of BIM on the supply chain as well as updating existing competencies and skills accordingly.

The Construction industry is historically known to be highly disintegrated and often depicted as involving a culture of "adversarial relationships" and "risk avoidance", intensified by a "linear workflow". As a direct consequence, this leads to a decrease in efficiency, punctuality, and efficient use of resources (Rezgui and Miles 2010, 2011; Alreshidi et al. 2018). Given the forecast of growth in the global construction market of over 70% by 2025 (Robinson 2013), the industry is faced with the challenge and opportunity to reduce energy demand, improve process efficiency, and reduce carbon emissions in line with the Energy Performance of Buildings Directive (2010/31/EU) (Li et al. 2019). In this context, energy efficiency demands adapted technology solutions, strategies (including training and education), and policy-making approaches which should be embraced by the entire supply chain across the whole life cycle of a project. The training and education landscape in construction exhibits the following characteristics (Palm and Thollander 2010; Chai and Yeo 2012; Alhamami et al. 2020):

- Numerous points of entry, a set of credentials, a broad range of training quality, and a diversity of funding opportunities;
- Fall in apprenticeship completions due to difficult economic conditions;
- Use of a more flexible self-employed labour pool due to unpredictability in the market;
- Low training and development activity, driven by the numerous self-employed tradesmen who often face an "earn or learn" impasse;

- The transitory presence of workers and the growing demand for training in the market deterring employers from investing in staff training;
- Lack of career planning and the tendency to adopt a supplier as opposed to a demand-driven model;
- Lack of a strategy of exercising Continuing Professional Development (CPD) and Continuing Craft Development across the industry.

Moreover, these characteristics are exacerbated by the fact that energy use and efficiency measures tend to focus primarily on the deployment of efficient technologies, such as highly energy-efficient construction products (e.g. facades) as well as renewable technologies, but less on energy management best practices, including training and education (Petri and Rezgui 2020). In fact, investments in technology, upgrading equipment, the introduction/adaptation of incentives, and strengthening of the regulatory frameworks generate improved efficiencies, but, without adapted training, the efficiency potential will not be attained (Backlund et al. 2012; Alhamami et al. 2020). In this context, the knowledge of built assets captured through digital models—moving from many data sites to IFC (Industry Foundation Classes)-based (or proprietary) centralised BIM servers—can offer invaluable context to enhance procurement, construction, and operation processes as well as paving the way to BIM-enabled, connected, and autonomous systems. The addition of the IoT, networks formed by physical devices embedded with circuits, software, sensors, actuators, and connectivity which allow these devices to interact and exchange data, creates room for more effective real-world incorporation into computer systems. The IoT is one avenue aiming to offer new ways of rethinking and resituating built environment data, with the potential to progress from an information delivery focus to value creation through an informed decision support perspective. The delivery of a building or infrastructure project involves a broad range of professions, from blue to white collars, which are now required to upgrade their skills in order to face this digital transformation of the industry. Such professions and roles are changing from traditional competencies which were required in a construction project to more information-driven skills promoted by the digitisation of the industry. Such skills and roles are continuously changing from one project to another especially since the gradual embodiment of BIM in design, construction, and operation processes alongside the quest to address energy efficiency (Petri et al. 2017, 2018; Hodorog et al. 2019). Therefore, the research problem which needs to be addressed involves the understanding of such dynamics in the Construction industry and associated implications for roles and skills with a view to harmonising as well as devising training and educational programmes for the current as well as the next generation of construction professionals. Such emerging industry requirements and trends can be captured with text-mining and natural language processing techniques applied on crawl data repositories such as social media content. Social media data (such as Twitter) are easy to access and embed a high value of semantics as online communities of users tend to openly express facts, statements, and requirements with high adoption rates. This paper leverages the accessibility and value of such open social media data and communities as well as their ability to report realities and trends which cannot be captured via traditional data sources and analysis (i.e. plain documents) in the attempt to capture BIM requirements, skills, roles, and competencies which can accelerate the digitalisation of the Construction industry (Boje et al. 2020).

Practitioners in the Construction industry are faced with an increasingly stringent and demanding regulatory landscape to decarbonise our built environment. BIM plays a central role in this strategic agenda. However, there is no established BIM training protocol that is aligned with the ever-changing demands and evolution of the industry. BIM training is currently static in that there is no systematic approach to review, adapt, and update existing BIM training, taking into account practitioners' ICT maturity and the BIM continuous evolution. To address this gap, we propose a combined analysis of secondary sources of evidence, such as scientific publications with social media information, using TF-IDF (Term Frequency-Inverse Document Frequency), association, correlation, and cluster analysis to infer roles and responsibilities as to the adoption of BIM in the industry. This forms a clear contribution to the body of knowledge in the field of BIM training and education with a view to promote sustainable interventions.

In this paper, we propose a methodology for identifying the frequency and correlation of roles and skills required for the implementation of BIM (Building Information Modelling) for energy efficiency. We use data mining techniques on a repository of BIM-related documents and Twitter records with the aim of capturing emerging BIM roles and associated skills in the field of construction. This involves parsing the collected datasets to identify associated terms, using the TF-IDF algorithm, and a bespoke importance formula based on metric cluster techniques. We have utilised a semantic BIM training portal (*energy-bim.com*) to conduct the analysis, which featured mechanisms for content aggregation from different BIM-related source materials. The outcomes of our research contribute to the transformation of the current traditional Construction industry towards a more sustainable and competitive industry and ultimately contribute to a higher quality of life in society. The Introduction is continued by "Related work" section, which outlines a review of the literature relevant to the research topic. "Overarching

methodology" section presents the underpinning methodology of the research. "Proposed text-mining approach" section illustrates the proposed text-mining approach. "Evaluation" section interprets and discusses the results, including key insights arising from the evaluation process. "Conclusions" section provides concluding remarks and explores possible avenues for future research.

## Related work

There is an abundance of literature exploring the application of text-mining techniques in a wide range of domains, including policy making (Apte et al. 1998; Ngai and Lee 2016; Antons et al. 2020). Chai et al. (2013) have demonstrated that the most significant advantage posed by the integration of text-mining processes into the policy-making process is to alleviate the problem of bounded rationality, a long-established policy problem (Meng 2009; Antons et al. 2020), described as a core principle for incremental policy change (Dror 1964). As text-mining applications are based on the adaptation principle, they create time-efficient processes in terms of data collection and information processing. This increases the time allowance of decision-makers to adapt to the actual challenging environment and produce better estimates of results. A significant example is represented by the early-warning signals that real-time text-mining applications can issue, such as predicting earthquakes from social media content (Sakaki et al. 2012).

Gao and Eldin (2014) developed a methodology for extracting text to locate employment credentials information relevant to the construction workforce professions, such as knowledge areas, skills, and expertise from employment information publicly available in online sources. The authors utilised search engines, which were market novelties at the time, to understand the job market and expectations of employers, while developing a system algorithm through statistically valuable pattern extraction, performed automatically within the qualifications selected previously, and then using them to detect the presence of such qualifications on new pages. Once the qualifications were determined, the LDA (Latent Dirichlet Allocation) model was used to identify skill groups which employers require. After aggregating the most relevant 10 keywords from each of the most relevant 10 topics, the authors concluded that adding more industry skills to the existing curriculum has the potential to help the next generations of students secure their career path in the Construction industry more firmly.

Forman et al. (2006) attempted to minimise the human effort in quantifying the issues in call logs from customer-support centres, arising from unstructured free-text fields. This research attempted to ascertain an accurate quantification aggregating expenditure broken down by the type of problem addressed in order to optimise human resource allocation through appropriately targeting the right engineers, providing the most effective means of identification, assistance, and recording of the most frequent problems. The requirement of the manual classification of calls is eliminated through underlying approaches such as new methods of text clustering, machine learning through the training of categorisers in an interactive manner, and quantifying categories.

Singh et al. (2007) have done longitudinal research over the course of a decade to determine the progress in hospitality human resource management and identify emerging trends. The data were collected and analysed by a text-mining algorithm in conjunction with human judgements. Both the results from the content analysed by the computer and the content generated through human judgement were integrated and conceptually graphed into a map readable by both computers and humans. This research resulted in nine major human resources management research topics, later interpreted based on the relevant timestamp and geographical region.

Succar and Sher (2014) have analysed the way in which organisational and educational institutions have started to adapt their delivery systems to meet changing demands on the market. This was one of the first papers to introduce taxonomies and conceptual models to clarify the mechanisms for filtering, classification, and aggregation of individual responsibilities into a competences database. It also discussed the benefits that the competency-based approach brings both to the industry and academia. In fact, according to the authors, individual BIM skills are the personal traits, technical skills, and professional skills required by an individual to conduct a BIM activity successfully or to produce a result linked to BIM. These skills, results, or operations could be measured against performance criteria and obtained by growth, training, and education or even improved by development.

Wu and Issa (2014) advise preparation as a way to boost BIM curving while acknowledging that recent graduates' qualifications are not sufficient to meet the demand of industry jobs. Rather, they propose that BIM education should train graduates to be ready to the degree that their BIM skills can be influenced by organisations according to their bespoke needs.

Meziane and Rezgui (2004) and Antons et al. (2020) emphasise the expertise of BIM for managing buildings and how a team leader with strong BIM skills can have a significant impact on project success and teamwork. The Construction industry enjoys recruiting future employees, not only possessing BIM technology experience, but also broad analytical knowledge.

Other social media mining algorithms which have been explored by numerous researchers such as Lopez-Castroman et al. (2019) and Song et al. (2018) are mainly focused on suicide prevention as well as crime forecasting and alert systems. We can observe that there is a gap present in the

current literature, represented by a lack of use of social media mining, with the scope of improvement for the BIM industry, particularly in regard to BIM roles and skills. Nevertheless, we could find numerous attempts to use the BIM industry field as an application for social media mining, such as the one of Zhang and Ashuri (2018) which attempted to mine the BIM design logs to discover connections between social network features and the production success of designers. Additionally, Kassem et al. (2018) attempted to identify the key competencies of the BIM expert positions which are selected on the basis of their quotes and the review of their skills overlap.

Barison and Santos (2011) undertook extensive research on the BIM roles and skills advertised in job ad descriptions, while also conducting a comparative analysis of the market demands and the efforts of universities and other institutions to integrate BIM into the curriculum. He also highlighted the importance of the correlation between BIM roles and skills for the companies which adopt management competency models.

This paper leverages on the notion of text and social media mining, including material from texts and web content, to comprehend BIM's dependence on energy-related ideas mainly linked to roles and skills, while presenting a methodology for classifying their frequency and correlations. The analyses are delivered by a semantic web platform (*energy-bim.com*), supporting TF-IDF techniques and importance evaluation methods to identify BIM skills and roles with corresponding dependencies.

Based on the critical review of the related literature, there is a clear gap in which no study has combined the analysis of secondary sources of evidence with social media information, using TF-IDF, association, correlation, and cluster analysis to infer roles and responsibilities as to the adoption of BIM in the industry.

## Overarching methodology

The aim of this research is to understand the implications of the digital transformation of the Construction industry, including the gradual adoption of BIM, in terms of required roles and skills to harmonise and devise training and educational programmes for the next generation of construction professionals.

The research utilises a positivist philosophical stance using a mixed-method consisting of quantitative and qualitative approaches to analyse secondary sources of evidence drawn from the academic and industry literature, as well as social media sources associated with authoritative institutions and experienced practitioners. As such, the research addresses the following two research questions:

1. How can we infer the roles and skills required as a result of the introduction of BIM in the Construction industry;
2. How to ensure that these roles, competencies, and skills are kept up to date as BIM is being widely deployed in the industry.

Informed by the consultation studies conducted as part of the EU H2020 BIMEET project (Petri et al. 2017), we have adapted and exposed a semantic web platform, *energy-bim.com*, which can manage, store, and analyse BIM-related information. The platform supports BIM knowledge sharing and enrichment within a community of BIM professionals and resources with a view to advance the implementation of BIM for energy efficiency in the construction sector. The *energy-bim.com* platform integrates various BIM-related data sources, based on which a set of TF-IDF and metric cluster methods are applied to determine relevant roles and skills around BIM.

### Harvesting BIM data from web sources

In the form of interactive, responsive, and user-oriented applications, which utilise the latest technologies, the *energy-bim.com* portal provides access to integrated BIM resources for a community of users and professionals. The framework is an open, scalable, polymorphic context-based solution with modules which allow BIM information and knowledge to be unlocked through a symbiosis of technology (Petri et al. 2014).

Figure 1 presents an overview of the *energy-bim.com* portal, by highlighting its primary widgets (Building Information Modelling, Sustainable Sourcing, Training, Sustainable Design Tools, Web Search, and Professional Networking), in conjunction with the refined search facility that filtrates the search results of the queries directed to the back-end ontology.

The platform retrieves and stores a repository of BIM data sources, including:

1. Documents such as scientific publications, standards, and regulations in the field of BIM for energy efficiency;
2. Twitter data from professional organisations, bodies, and actors working in the field of BIM and energy efficiency

We have retrieved around 80 key BIM-related publications and regulations alongside a total of 4 million tweets based on which the analysis has been conducted. We have developed a search service which searches the BIM knowledge base from several authoritative URIs (Uniform Resource Identifiers) as part of the application. The BIM query submitted contains a number of related ontological principles to improve the accuracy and retrieve the returned data. The search service also includes data from a number of reputable

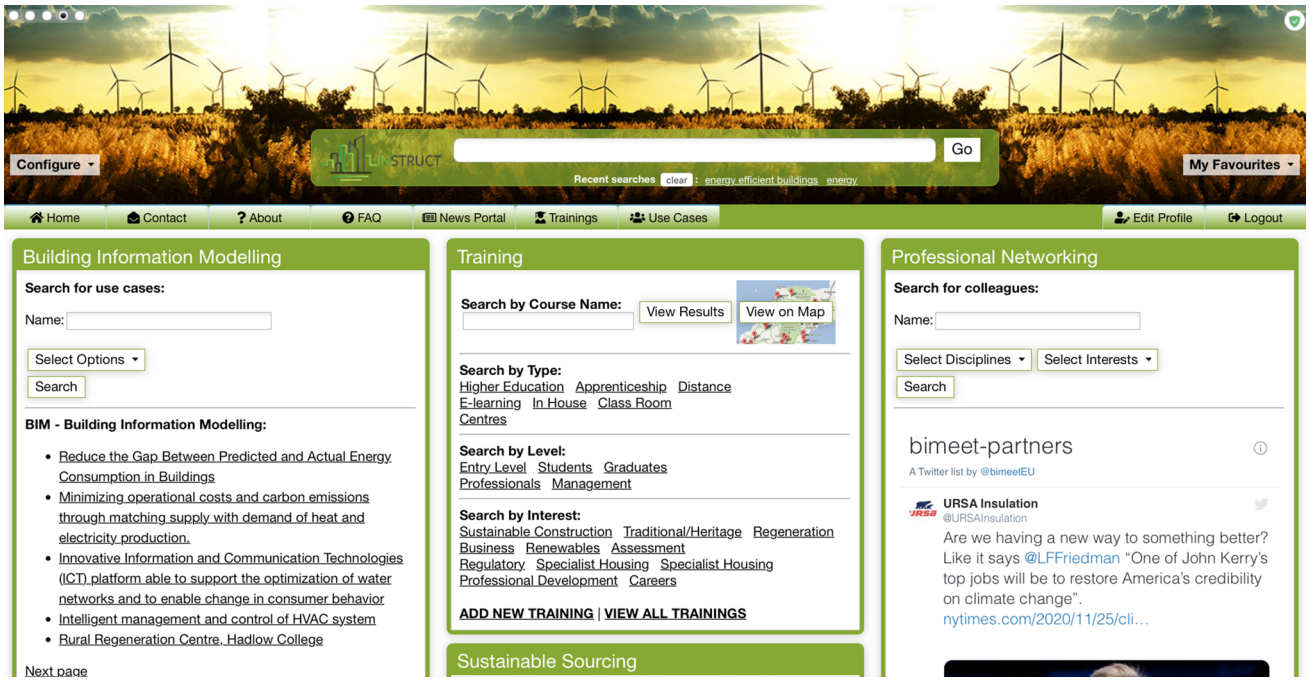


**Fig. 1** The energy-bim.com portal

BIM-related sources via the web. These sources can be suggested by users and tested in terms of energy efficiency by a group of experts.

The scalable, multi-focal, context-based widgets of the *energy-bim.com* platform are the creative aspect which can be reconfigured to answer evolving user contexts and (BIM-related) questions while allowing for smart BIMxt (Building Information Modelling Extension Network) information and the exploration of knowledge. Every service has a related toolbar, which users can update and control remotely. Analysing usage data and feedback has helped to define several key problems to address in future product launches.

## Conducting semantics analysis of the BIM knowledge-based data sources

The ontology running in the background of the *energy-bim.com* portal has facilitated the mining and analysis phases of the BIM knowledge harvesting process (Petri et al. 2014). The query and extension methods within the framework provide the primary use of an ontology to drive the search engine. Firstly, the terms in the ontology are used when entering search terms (using the query method) to give keyword suggestions. Secondly, the relationships between words are used to help users expand/limit their questions on the basis of ontological suggestions.

The ontology service is built upon a basis of semantic vectors to ensure the required level of knowledge for the platform is met. The ontology seeks to enhance and extend the contents and current domain requirements with additional concepts and aspects taken from: (i) an engineering-specific knowledge repository and (ii) information structures which are the basis for calculations, simulations, and resources for monitoring compliance. The ontology uses term frequency—inverse document frequency (TF-IDF) and metric cluster algorithms to detect related ontological concepts in and around a knowledge-based repository. We measure the degree of significance (semantic) in more detail for each definition and facet of the text as well as the entire collected documentary repository. To determine the level of importance associated with relationships between concepts, a process is implemented which specifies the number of co-occurrences of concepts in the document. This clustering algorithm calculates the difference between two terms in the measurement of its correlation factor.

The *energy-bim.com* platform has been developed based on preliminary consultations and user requirements, which have indicated that they require access to BIM knowledge and training resources. We have incorporated scientific publications and regulations alongside Twitter data in the attempt to create a scalable knowledge-based environment accessible to users and actors from the Construction industry. The platform has been utilised to support and deliver the

analysis and the entire methodology presented in "Proposed text-mining approach" section.

## Proposed text-mining approach

The methodology comprises a mixed content mining and analysis research method to determine BIM roles, skills, and competencies for the Construction industry. We undertook an incremental research process from the content aggregation to data curation and analysis intended to extract knowledge related to BIM roles and associated skills for energy efficiency.

We have conducted knowledge mining on two data repositories hosted by the *energy-bim.com* platform for two purposes: (i) one identifies scientific publications, data, and regulations, and (ii) the second repository is comprised of a collection of tweets from social media sources.

The steps covered by our methodology are broadly illustrated in Fig. 2. In order to confirm the roles and competencies required by this training process, a scientific literature repository comprised of 80 BIM publications and related analytics has been implemented using the steps below.

The publications repository was created after searching popular scientific portals such as Google Scholar and Web of Science using specific keywords to filter relevant results to BIM for energy efficiency. A comprehensive literature search based on several matching criteria for the title, abstract, and keywords for papers was first conducted through the Scopus search engine for scholarly publications. We have used several keywords to conduct these such as "building information modelling", "energy efficiency", "role", "skill", "training", "education", and "competencies". This review identified articles, published between 2009 and 2019, which have been imported into the *energy-bim.com* platform as a knowledge-based repository and used in the analysis phase. The fairly recent time interval is justified by the fact that BIM for energy efficiency is a novelty in terms of technological advancements.

The social media repository was created using the *energy-bim.com* platform which fetched a corpus of 40 million tweets from a selection of Twitter accounts belonging to the friends and followers of the most prominent users of the *energy-bim.com* platform (illustrated in Table 1). In the filtering stage, the tweets were filtered using regular expressions and SQL queries, thus obtaining a corpus of 60,000 selected tweets for further analysis. The association & importance stage determines the pairs of roles and skills while computing an importance score for each role and skill. At the final stage (correlation), TF-IDF and metric cluster methods are applied to establish concept dependencies for roles and skills. This has resulted in an ontology of concepts of roles and skills, together with their associated correlation factors. Below we explain the entire data collection and analysis process supported via the *energy-bim.com* platform.

### Data collection

We use a generic mining function $f(t)$ $f(t) : C_t \rightarrow \{R, S\}$, where $C_t = \{t_{e1}, t_{e2}, \ldots t_{em}\}$ is a collection of endpoints (Twitter accounts) and each $t_{ei}$ is a twitter account which generates a set of tweets, $R = \{r_1, r_2, \ldots r_b\}$ is a set of retrieved roles, and $S = \{s_1, s_2, \ldots s_c\}$ is a set of retrieved skills.

During the collection process, each tweet has to be pre-processed in order to make it readable for our parsing algorithms and regular expressions. In a scientific publication, a sentence with BIM-related keywords is the equivalent of a tweet. Additionally, this process increases the reliability of the results and the reduction of bias (for example, a role being present as part of a camelCase construction).

Table 1 presents the portfolio of companies and organisations which have been used as primary data sources for social media mining.

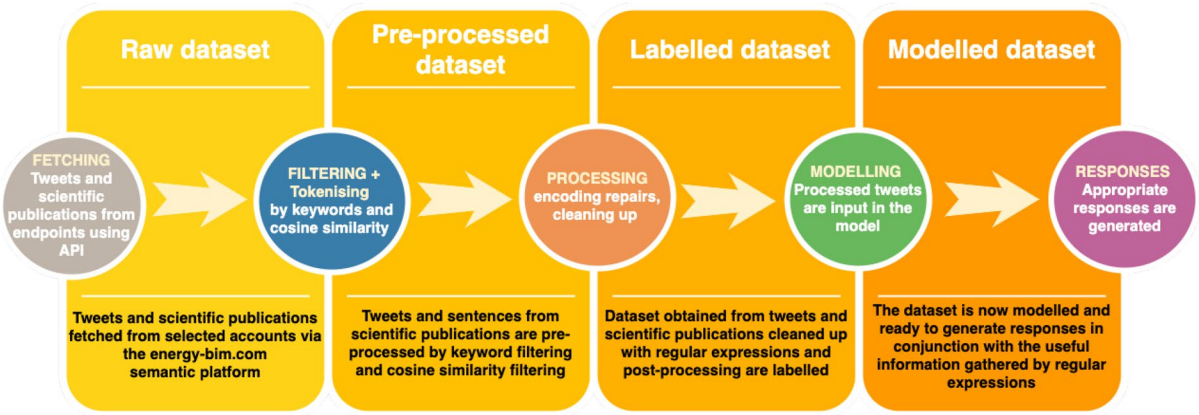


**Fig. 2** The steps of the process involved in our methodology

**Table 1** List of companies/ organisations used as primary data sources and their Twitter account

| Name of organisation | Twitter account |
|---|---|
| Group CSI | https://twitter.com/groupecesi |
| INESs Solaires | https://twitter.com/inessolaire |
| BRE Academy | https://twitter.com/BREAcademy |
| Écoles des Ponts Paris Tech | https://twitter.com/EcoledesPonts |
| ESTP | https://twitter.com/estpparis |
| Universite de Lièege | https://twitter.com/UniversiteLiege |
| UC Louvain | https://twitter.com/UCLouvain be |
| Citt'a di Modena | https://twitter.com/cittadimodena |
| ORSYS Formation | https://twitter.com/ORSYS |
| BEC partners SA | https://twitter.com/becpartners |
| Middlesex University | https://twitter.com/MiddlesexUni |
| House of Training | https://twitter.com/Houseoftraining |
| Sapienza Universita | https://twitter.com/sapienzaRoma |
| Scuola Master F.lli Pesenti, Politecnico di Milano | https://twitter.com/master pesenti |
| Le Moniteur | https://twitter.com/Le Moniteur |
| Technical University of Denmark | https://twitter.com/DTUtweet |
| Norwegian University of Science and Technology | https://twitter.com/ntnu |
| UIC Barcelona | https://twitter.com/UICbarcelona |
| Mensch und Maschine | https://twitter.com/MuMDACH |
| Zigurat | https://twitter.com/Ziguratdigital |
| BIMEET EU | https://twitter.com/bimeetEU |
| H2020EE | https://twitter.com/H2020EE |
| H2020 BIM plement | https://twitter.com/H2020BIMplement |
| ECTP Secretariat | https://twitter.com/ECTPSecretariat |

The capturing process for the tweets and scientific publications was supported by a list of organisations identified from three sources:

1. IP detection and organisational identification forensic algorithms;
2. Followers of the "@BIMEET" Twitter account;
3. BIM institutions indicated by partners.

We considered these authoritative organisations to be highly relevant as driving forces of the continuous transformation of the BIM industry as they are active users of the *energybim.com* platform. We have utilised the 60,000 filtered tweets from the total corpus of 4 million posted by the portfolio of companies presented in Table 1.

## Data filtering

For the filtering process, we use a set of expressions, $N_e = \{exp_1, exp_2, ...exp_n\}$, where $exp_i$ is an expression as defined in the list of expressions presented below, filtering $C_t''$, $C_t'' \subset C_{t_i}$, where custom tuples of roles and skills are defined as $C_t'' = \{g(t_{eij}), \rightarrow \{r_i, s_j\} \in \mathbb{N}_E\}$

The following set of regular expressions (RegExp) and pattern matching rules were needed to filter information, due to the noisy nature of social media in which casual spelling, grammar, and brief expressions are often used. The same expressions were also used to filter data found in scientific publications, a process further aided by manual recognition.

```
+((contractor\manager\designer\engineer\client\)+\skills+\).+ (\energy+ \construction)
+((\BIM\construction\energy)+\skills+\).+ (\need+ \require)
+((\BIM\construction\energy)+\roles+\).+ (\need+ \require)
+((\BIM\construction\energy)+\actors+\).+ (\skills+ \competencies)
+((\BIM\construction\energy)+\knowledge+\).+ (\requirements+ \require)
+((\BIM\construction\energy)+\skills+\).+ (\need+ \require)
+((\BIM\construction\energy)+\competencies+\).+ (\need+ \require)
+((\skills\competencies\knowledge\expertise)+\BIM+\).+ (\energy+ \construction)
```

As representative starting points, the concepts of "skills" and "roles" were created for the regular expressions with the primary objective to filter the inaccurate context before analysis. Roles were associated frequently with keywords including "construction", "skills", and "energy". Several terms, including "training" and "knowledge", could be found in sentences encompassing both roles and skills.

## Establishing association, importance, and correlation

To determine the relationships between roles and skills, the following functions are also used for determining the associations, correlations, and importance between roles and skills:

(a) $f(a)$ for determining the association between roles and skills facilitated by TF-IDF techniques (as described in Eq. 1): $f(a) : [R, S] \rightarrow [RS'_a, A]$, where $RS'_a$ is a set of roles and skills pairs with a Boolean value $A$ that states if the role and skill pair are associated or not.
(b) $f(i)$ for determining the importance (as described in Eqs. 2 and 3): $f(i) : [R, S] \rightarrow [RS'_c, I]$, where $RS'_i$ is a set of roles and skills pairs with their importance value $I$.
(c) $f(c)$ for determining the correlation through the proximity concept (as described in Eqs. 4 and 5): $f(c) : [R, S] \rightarrow [RS'_c, F]$, where $RS'_c$ is a set of roles and skills pairs with their correlation value $F$.

## Objectives

### Frequency determination

Our first objective was to determine the frequency of each BIM role in the corpus of 60,000 tweets and sentences in the 80 scientific publications. Conversely, our second objective was to determine the frequency of each BIM skill in the filtered corpus of data. The frequency of a role or skill has been determined based on the formula below:

$$TF(t) = \frac{frequency\ of\ t}{total\ terms\ of\ the\ same\ type\ as\ t} \quad (1)$$

From Eq. 1, $t$ is either a skill or a role and the *type of* $t$ can be either a skill or a role. This has facilitated the classification of skills and roles with a view to understanding their importance in the overall dataset.

### Association determination

Our third objective was to determine the associations between the BIM roles and skills with their associated importance.

Figure 3 illustrates the algorithm we used for determining the roles and skills which are associated in our dataset. We took the raw database and the list of previously determined roles and skills as an input. We then created two combination matrices. Each element in *roleSkillCombinationCountMatrix* has been used in Eq. 2, and each element in *skillRoleCombinationCountMatrix* has been used in Eq. 3. In the context of scientific publications, we considered a role and a skill to be associated if they could be found in the same sentence.

### Importance determination

Our fourth objective was to determine the importance of each role and skill, taking into account the averaged number

```
1: userList = SELECT DISTINCT FROM users
2: INITIALISE combinationList, roleSkillCombinationMatrix, skillRoleCombinationMatrix
3: for each user in userList do
4:     INITIALISE currentUserRoleList
5:     for each role, skill do
6:         ADD role TO currentUserRoleList
7:     end for
8:     INITIALISE currentUserSkillList
9:     for each skill do
10:        ADD skill TO currentUserSkillList
11:    end for
12: end for
13: for each combination in combinationList do
14:    for each role, skill do
15:        if role, skills IN combination then
16:            ADD role, skill TO roleSkillCombinationCountMatrix
17:            roleSkillCombinationCountMatrix[role][skill from combination] += 1
18:        end if
19:    end for
20: end for
```

**Fig. 3** Roles and skills correlation + combination matrix algorithm

of occurrences in conjunction with all its counterparts as in Eq. 2.

$$\text{Importance(skill)} = \frac{1}{n}\sum_{n=1}^{n}(\text{no. of occurrences(association(skill + role))} \tag{2}$$

where $n$ is the number of different associations between a skill and different roles.

$$\text{Importance}(role) = \frac{1}{n}\sum_{n=1}^{n}(\text{no. of occurrences(association(role + skill))} \tag{3}$$

where $n$ is the number of different associations between a role and different skills.

#### Correlation determination

Our final objective was to determine the degree of correlation between the terms (roles and skills) occurring within the context of the same tweet or sentence. We did this by determining semantic relationships between concepts, through applying the metric clusters method (Baeza-Yates et al. 1999) which factors the number of co-occurrences of concepts with their proximity in the text. To determine ontological dependencies between concepts, we have applied a combination of TF-IDF and metric clusters methods to infer the correlation factor between BIM roles and skills (Rezgui 2007).

$$C_{\mathrm{u,v}} = \sum_{k_\mathrm{i}\in V(S_\mathrm{u})}\sum_{k_\mathrm{j}\in V(S_\mathrm{v})}\frac{1}{r(k_\mathrm{i},k_\mathrm{j})} \tag{4}$$

$$C_{\mathrm{u,v}} = \frac{1}{\mathrm{Min}\left[r(k_\mathrm{u},k_\mathrm{v})\right]} \tag{5}$$

The correlation technique is presented in Eq. 4 where the distance $r(k_i, k_j)$ between two keywords $k_i$ and $k_j$ is calculated as the number of words interjected between two terms in the same tweet or sentence. $V(S_u)$ and $V(S_v)$ represent the tuples of keywords which have the stems $S_u$ and $S_v$ linked with them. To improve the degree of generalisation, the application instances of this formula in our experiments discarded different semantic discrepancies of concepts within the Twitter text and instead used a more efficient correlation technique as in Eq. 5, where $r(k_u, k_v)$ is a depiction of the minimum number of words which separate the two keywords $k_u$ and $k_v$ in each individual tweet (or sentence).

The hypothesis when applying Eq. 4 is that a low difference between the correlation factor and the value of 1 for the denominator is a representation of a high strength of correlation between the two terms.

In order to establish a threshold for the correlation factor, we decided to limit the application of the formula to the instances of tweets which embed both of the two terms, while specifying the following exception: should the minimum distance between two analysed keywords be null, the default correlation factor will still be "1". On the occurrence of this exception, the correlated terms are considered as candidates to be part of a new, unified term.

## Evaluation

In this section, we report the outcome of the analysis involving frequency, association, and importance of roles and skills based on the methodology presented in the previous section.

#### BIM roles frequency

In this experiment, we seek to analyse roles based on their frequency in the document corpus. Figure 4 presents a classification of roles based on their frequency, where roles such as "energy modeller" and "energy procurement specialist" have been identified as important based on the analysis. A lower frequency has been attached with roles such as "HR specialist", "finance specialist", and "communication officer", demonstrating a key area of interest for developing more BIM competencies and educational programmes.

From Fig. 4, it can also be observed that BIM has an impact on different areas of practice ranging from "clerical" roles to "scientists", which brings a relatively high diversity of BIM roles.

We can observe that the scientific publications repository embeds a greater variety of roles than the social media corpus due to the presence of lower-skilled roles such as "plasterer", "dry liner", and "bricklayer". However, as expected from the difference in the size of the repositories, the frequency of the roles found in scientific publications is lower. Nevertheless, we can observe the mutual presence of several high-skilled roles such as "BIM teacher", "facility manager", "building professional", and "quality assurance manager".

From the Construction industry perspective, this experiment demonstrates that disciplines and roles need to adapt

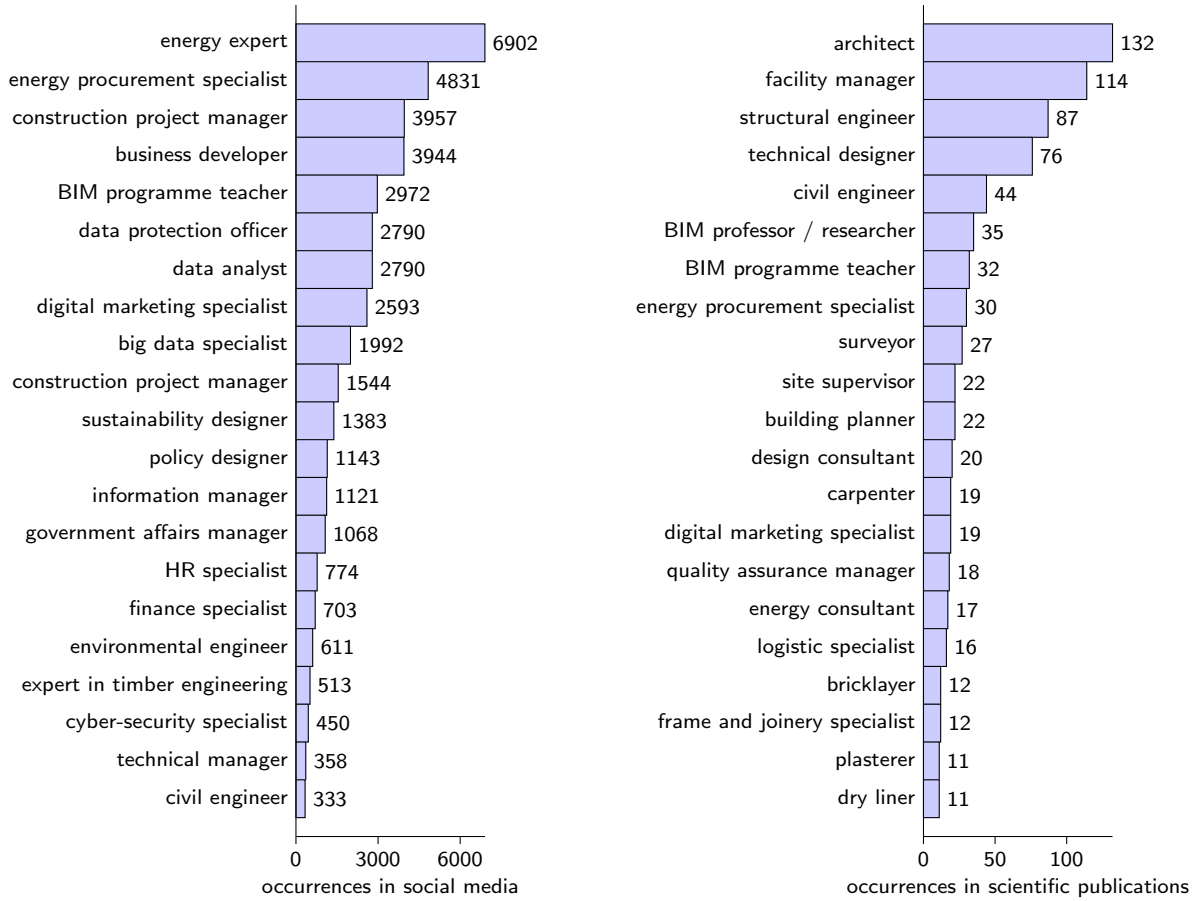


**Fig. 4** Frequency of BIM roles on social media (left) vs. scientific publications (right)

to the emerging challenges identified in the industry, with a key emphasis on "energy"-related roles and competencies which present the highest frequency in the analysis.

## BIM skills frequency

In this experiment, we identify skills in relation to their frequency. Figure 5 presents a classification of skills based on their frequency, where skills such as "4D simulation", "construction", and "urban planning" have been identified as important from the reports in users' tweets. A lower frequency has been attached to skills such as "structural engineering software" and "structural analysis".

From Fig. 5, it can be identified that BIM has a direct competency relation to advanced skills such as "urban planning" and "4D simulation". We can observe the mutual presence of specific skills in both repositories such as "Revit operation", "coordination skills", and "energy management", while also noticing a more extensive variety of skills in scientific publications, ranging from management to highly specialised roles such as "solar and wind panel operation". As expected, due to the difference of repository size, we can notice a lower frequency of roles from scientific publications.

We can observe that the skills with the highest frequency are "4D simulation", "construction", and "urban planning". These are followed by "training skills", "leadership", and "management", all of which have a high frequency. The skills directly correlated with the "energy" work field are in the middle range of the spectrum. We can also observe skills related to *business & management* such as "marketing skills", "speaking skills", "entrepreneurial skills", and "cooperation skills". The results also report new skills related to *technology* such as "ICT", "IoT", "virtual reality operation", "Revit operation", "automation", and "analytics".

The experimental results also report on the necessity to advance the development of new skills such as "4D simulation" as well as a greater understanding of ICT and IoT advancements, along with their applicability for the Construction industry. These findings are showing an intrinsic transformation of the Construction industry and the

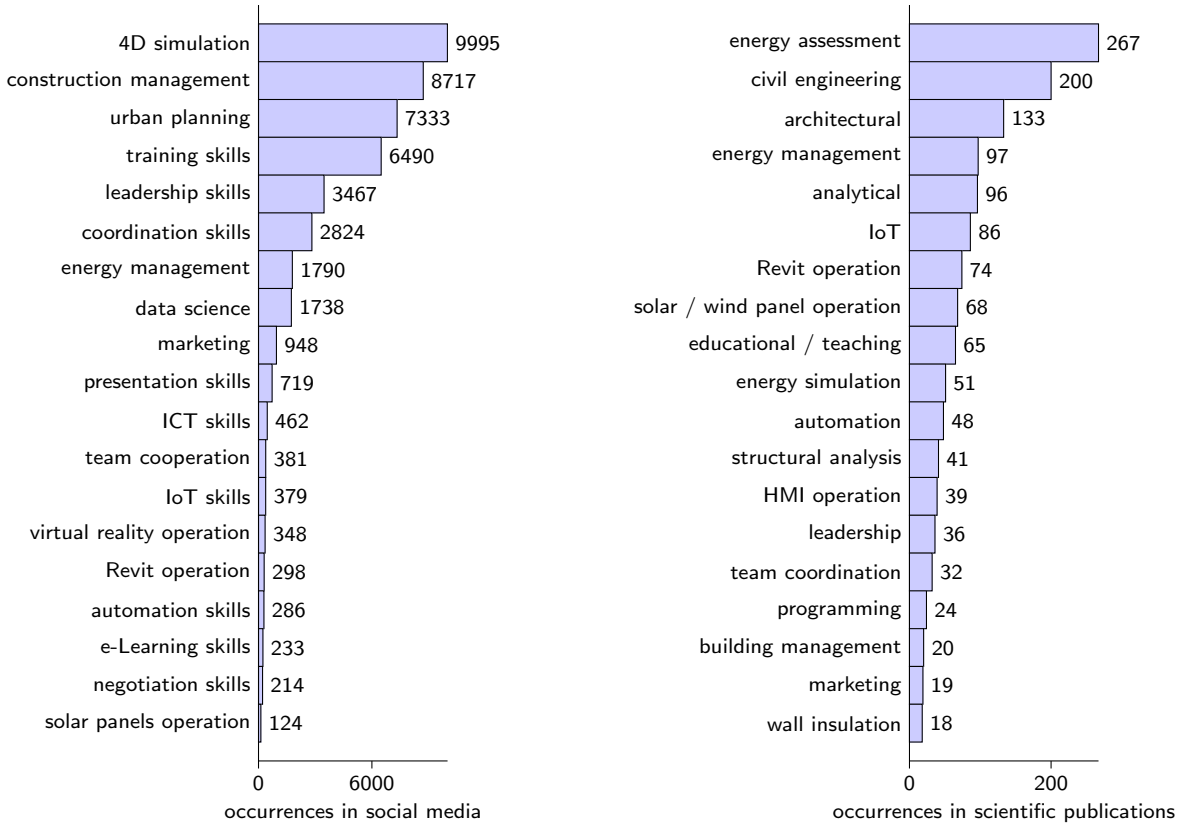


**Fig. 5** Frequency of BIM skills on social media (left) vs. scientific publications (right)

advantages of digitisation for supporting more informed and efficient energy practices.

## Importance of roles in relation to skills by category

The main objective of this experiment was to determine a correlation between the importance and the frequency of each role and skill. The importance and frequency are defined by the formulas presented in "Proposed text-mining approach" section . In Fig. 6, we have provided the outcomes of analysing the correlation between frequency and importance of roles and skills, with frequency on the *X*-axis and importance on the *Y*-axis. We can generally observe a positive relationship between the frequency and the importance.

For energy-specific roles, we can observe that the role having both the highest frequency and the highest importance is "energy engineering/expert", followed by "energy procurement specialist", while the "information manager" has a high importance but a low frequency, followed from here by "economy specialist".

With regard to management roles, the roles having both the highest frequency and the highest importance are "digital marketing specialist" and "construction project manager". The role which has a high frequency but low importance is "architecture & construction project manager". For research roles, the most important roles are "BIM professor/researcher", "BIM teaching assistant", and "BIM programme teacher".

In relation to technology roles, the roles with high importance are "data protection officer", "digital marketing specialist", and "big data specialist" with lower frequency for "user experience designer" and "technicians".

## Frequency and importance of skills in relation to roles

The resulting analysis depicts the correlation between frequency and importance in relation to BIM roles and skills, with frequency on the *X*-axis and importance on the *Y*-axis. From the results, we observe a linear trend with respect to the correlation between frequency and importance. Skills such as "coordination", "energy management", and "leadership" are characterised by high importance, but relatively low frequency. At the mid-point on the scale of both importance and frequency, we can notice skills such as "urban planning" and "training skills". We can observe skills with similar low frequency, such as "IoT", "cooperation", and

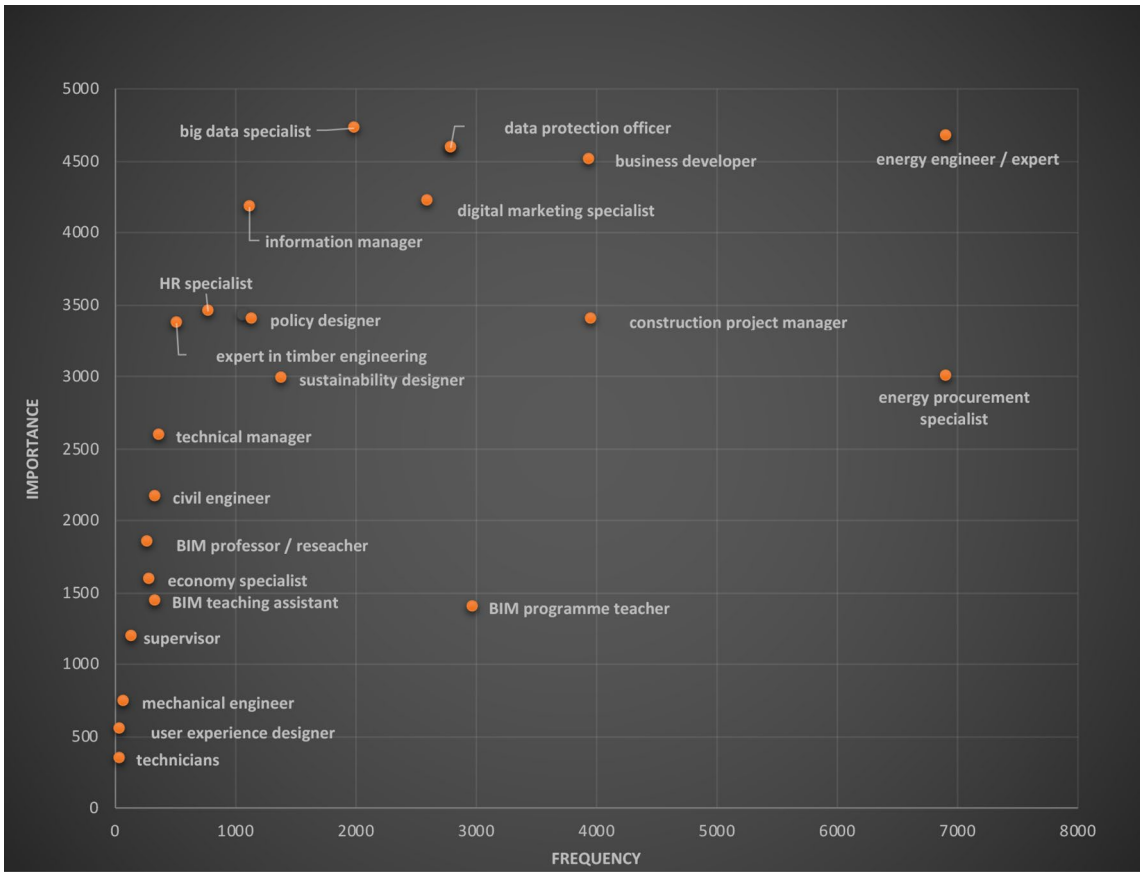


**Fig. 6** Frequency and importance of roles in relation to skills

"ICT". At the lowest end of the scale, we can notice skills with very low frequency such as "lifelong learning", "presentation skills", and "digitisation skills" (Fig. 7).

The results report numerous trends in relation to expertise and competencies required in the Construction industry. Such expertise can be obtained by developing a set of skills and roles around BIM and energy, supported by big data technologies to tackle digitisation and improve sustainable practices of the industry.

## Correlation factor between roles and skills

Further, we have determined the correlation factor between the top 100 combinations of roles and skills and converted the results into the format of an ontology, having the roles and skills as classes and the correlation factor as the object properties.

From Fig. 8, we can observe the presence of some central roles and skills with a large number of edges, such as "energy storage committee", "construction", "management", "energy transition", "energy performance", and "energy management". We can observe a cluster formed around the nodes containing the "energy" keyword, with "energy storage committee" being the most transited node.

We have also calculated a correlation factor determined according to Eq. 5. It can be observed that there is also a multitude of roles and skills which possess the maximum correlation factor of 1 (100%). These correlations illustrate a number of industry migration and transformation trends which identify the areas in the Construction industry where new training and education programmes are required.

The semantic analysis provides a clear illustration of the key concepts involved in promoting BIM for energy efficiency and their dependencies based on a correlation factor metric. Such analysis can facilitate a holistic representation of possible zones of improvements that need to be addressed in the construction sector. With the use of ontological conceptualisation, we have identified existing informational clusters that reflect various levels of digitalisation in the sector.

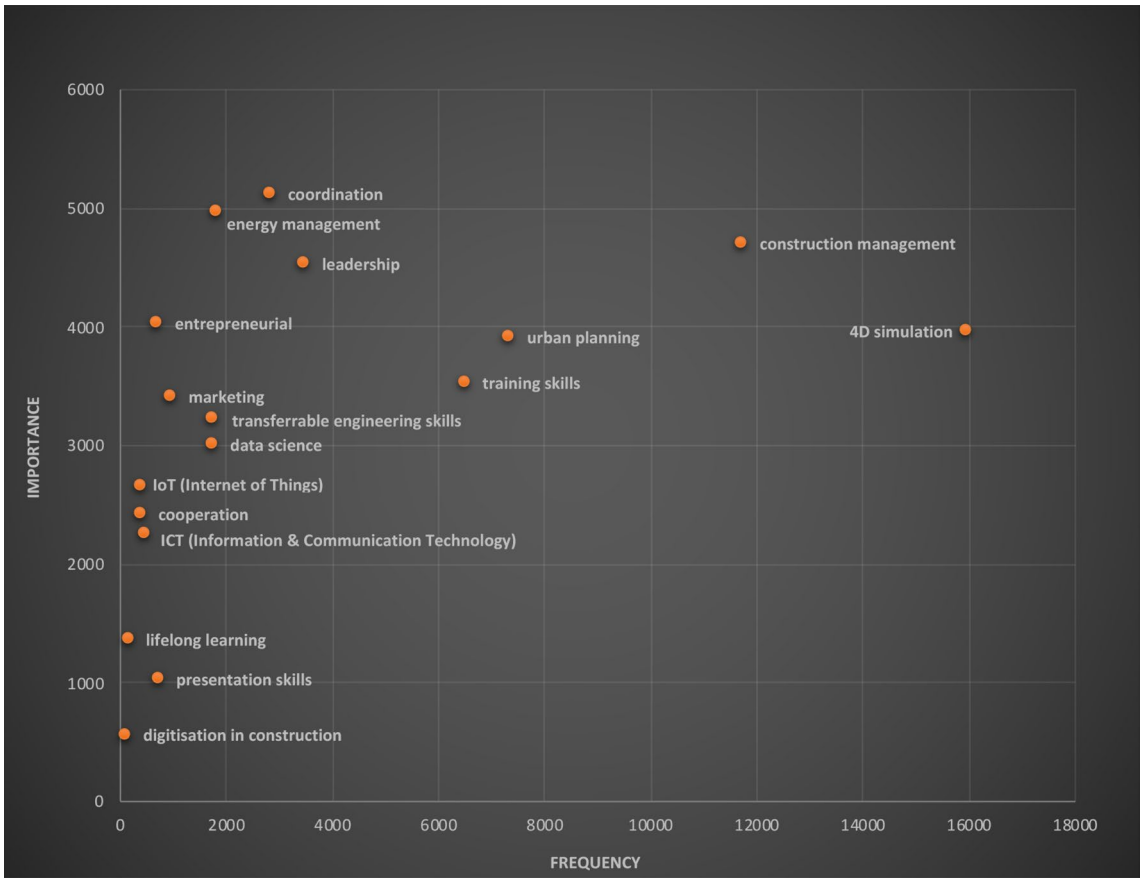


**Fig. 7** Frequency and importance of skills in relation to roles

## Key insights into the evaluation process

The proposed research provides encouraging results which have been validated in the context of the EU H2020 BIMEET project, in collaboration with our construction energy value chain, as evidenced on our web portal (*energy-bim.com*).

We have analysed a total of 40 million tweets sourced from a selection of Twitter accounts belonging to the friends and the followers of the most prominent users of the *energy-bim.com* platform, as illustrated in Table 1. This corpus was later filtered down to "60,000 tweets" of relevance, using regular expressions. In order to confirm the roles and competencies required by this training process, a scientific literature repository comprising "80 BIM publications" was analysed, mainly authored by the same organisations and connected authors.

Due to the restricted size of our dataset, the results might be influenced by a certain degree of bias, as they were derived from a random sample of organisations involved in the BIM industry, consisting only of the organisations using the *energy-bim.com* portal and their friends and followers. This analysis could be further scaled up for the entire BIM domain and for a wider representative sample of the Construction industry by considering partners and organisations from other BIM collaboration projects, multiple countries, and areas of expertise. This will have the potential to convey more information about training needs in the BIM industry.

The research could also benefit from the potential use of other text-mining techniques in addition to TF-IDF and cluster analysis. As evidenced recently by Singh (2016), techniques such as NLP (Natural Language Processing) and classification theoretical are useful for deriving useful knowledge for educational stakeholders.

As part of our EU H2020 BIMEET project, we have captured a database of training programmes provided by BIM accredited institutions that is actively updated, enabling institutions to continuously update their offerings (either scheduled or on-demand) with a detailed level of granularity. The training programmes are split into learning outcomes, prerequisites, and levels of expertise expected after completion. The findings of our research can be associated with the learning outcomes described by the organisations, and recommendations for improvement could also be derived. These results could be further disseminated to other stakeholders in order to encourage

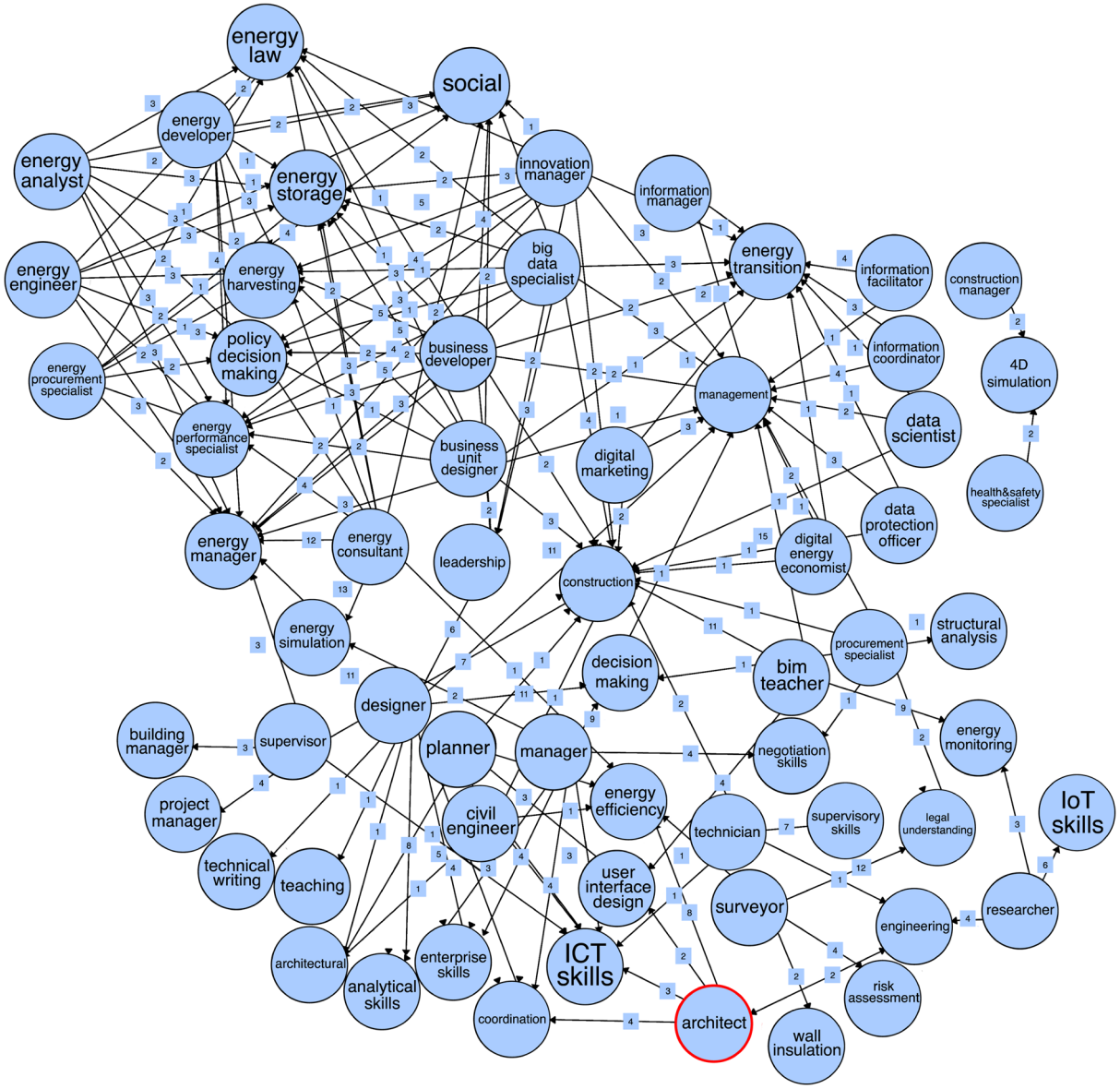


**Fig. 8** Ontology visualiser for correlation between BIM roles and skills

them to implement practical policies which take into account the newly found correlations between roles and skills.

We acknowledge our research does not incorporate emerging concepts in the industry, such as digital twins and scan to BIM. This is due to (a) the scope of the IFCs (Industry Foundation Classes) and (b) the sample used in the context of the research. Further, this limitation can be overcome by (a) leveraging semantic associations using our two data sources (social media and scientific publications) and (b) scaling up our sample to include other third-party social media sources drawn from other disciplines and countries.

The authors have recently initiated follow-on research in the context of the H2020 INSTRUCT project to address the training needs of blue- and white-collar staff across the life cycle of a construction project. This will give us the opportunity to stress-test the proposed text-mining approach, while benchmarking other text-mining algorithms.

## Conclusions

This paper elaborates on an in-depth analysis of primary sources of evidence for the identification of roles and associated skills necessary for improving the implementation of BIM for energy efficiency training in the construction market. This involved the analysis of a large corpus of

documents and Twitter data using text-mining techniques to extract BIM roles and skills as well as to infer their associations and linkages.

An approach was proposed to address the first research question consisting of four phases from the dataset collection and filtering to semanticisation and knowledge extraction. The approach presented in "Overarching methodology" section is delivered through the use of a semantic web platform which automates the update of the roles and skills required to reflect the changes of adoption and deployment of BIM in industry.

This study demonstrates that BIM is a complex undertaking involving multiple disciplines and domains from finance and management to engineering and technology with significant implications for research and education. Based on such findings, we identify key industry trends with an emphasis on competencies, skills, and roles required to implement more sustainable BIM practices for energy efficiency. Our research methods are generic and can be utilised in wider research contexts to address the ongoing climate change phenomenon.

The analysis has evidenced the importance of socio-organisational aspects in construction skills and the need to stimulate and sustain the advancement in the current and ongoing BIM implementation landscape. These findings corroborate existing literature, in which these BIM positions and associated skills are generally overlooked. For example, the majority of studies raised the awareness of an imminent need for vocational education to up-skill the construction workforce to address consistently and holistically the energy efficiency agenda. However, the Construction industry is renowned for its resistance to change which may hinder the current digitalisation agenda. As relevant correlation factors between BIM roles and skills were determined, it is worth highlighting that these correlations will change to reflect the continuous evolution of BIM roles and skills which demonstrate the incremental adoption, and improvement in maturity, of BIM in industry.

Also, the research has evidenced the role of BIM as an enabling technology for energy modelling. The use of BIM, i.e. the IFCs, falls short in describing complex energy systems. This is where the authors have augmented BIM with additional semantic resources, i.e. ontologies, in domains such as water, energy, and wider infrastructures, to represent complex artefacts. In fact, training for energy efficiency can be extended to include tools and software which makes use of BIM, augmented with additional concepts drawn from these ontologies to perform various forms of engineering analysis, including energy modelling. In addition to the acknowledged role of social media, a programme of change is necessary to fully embrace the digital transformation of the Construction industry. A holistic methodology, encompassing additional sources of information, was, therefore, required for the assessment of BIM with associated competencies and training programmes.

Future research involves using a project-based approach for evidencing the correlation between roles and skills by deploying the proposed methodology on real-world projects, thus providing tailored recommendations as to the set of skills required for each identified project role. As such, the authors will organise further consultations with the stakeholders involved in real-world projects and validate the assigned roles and skills, as well as the ensuing methodology, further informing the evolution of BIM training programmes delivered in the Construction industry.

**Acknowledgements** This work is part of the EU H2020 BIMEET project: "BIM-based EU-wide Standardised Qualification Framework for achieving Energy Efficiency Training" (grant reference: 753994).

**Data availability** Some or all data, models, or code which support the findings of this study are available from the corresponding author upon reasonable request.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.



## References

Alhamami A, Petri I, Rezgui Y, Kubicki S (2020) Promoting energy efficiency in the built environment through adapted bim training and education. Energies 13(9):2308

Alreshidi E, Mourshed M, Rezgui Y (2018) Requirements for cloud-based bim governance solutions to facilitate team collaboration in construction projects. Requir Eng 23(1):1–31

Antons D, Grünwald E, Cichy P, Salge TO (2020) The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. R&D Manag 50(3):329–351

Apte C, Damerau F, Weiss S et al (1998) Text mining with decision rules and decision trees. Citeseer. https://doi.org/10.1016/s0167-739x(97)00021-6

Backlund S, Thollander P, Palm J, Ottosson M (2012) Extending the energy efficiency gap. Energy Policy 51:392–396. https://doi.org/10.1016/j.enpol.2012.08.042

Baeza-Yates R, Ribeiro-Neto B, et al. (1999) Modern information retrieval, vol 463. ACM press New York

Barison M, Santos E (2011) The competencies of bim specialists: a comparative analysis of the literature review and job ad descriptions. In: Computing in Civil Engineering (2011), American Society of Civil Engineers Library, pp 594–602, https://doi.org/10.1061/41182(416)73

Boje C, Guerriero A, Kubicki S, Rezgui Y (2020) Towards a semantic construction digital twin: directions for future research. Autom Constr 114:103179

Chai KH, Yeo C (2012) Overcoming energy efficiency barriers through systems approach-a conceptual framework. Energy Policy 46:460–472. https://doi.org/10.1016/j.enpol.2012.04.012

Chai J, Liu JN, Ngai EW (2013) Application of decision-making techniques in supplier selection: a systematic review of literature. Expert Syst Appl 40(10):3872–3885. https://doi.org/10.1016/10.1016/j.eswa.2012.12.040

Dror Y (1964) Muddling through-" science" or inertia? Public administration review pp 153–157, https://doi.org/10.2307/973640

Forman G, Kirshenbaum E, Suermondt J (2006) Pragmatic text mining: minimizing human effort to quantify many issues in call logs. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 852–861, https://doi.org/10.1145/1150402.1150520

Gao L, Eldin N (2014) Employers' expectations: a probabilistic text mining model. Proc Eng 85:175–182. https://doi.org/10.1016/j.proeng.2014.10.542

Hodorog A, Alhamami AHS, Petri I, Rezgui Y, Kubicki S, Guerriro A (2019) Social media mining for bim skills and roles for energy efficiency. In: 2019 IEEE international conference on engineering, technology and innovation (ICE/ITMC), IEEE, pp 1–10, https://doi.org/10.1109/ICE.2019.8792571

Kassem M, Raoff A, Liyana N, Ouahrani D (2018) Identifying and analyzing bim specialist roles using a competency-based approach. In: Creative construction conference 2018, CCC 2018, 30 June–3 July 2018, Ljubljana, Slovenia, https://doi.org/10.3311/ccc2018-135

Li Y, Kubicki S, Guerriero A, Rezgui Y (2019) Review of building energy performance certification schemes towards future improvement. Renew Sustain Energy Rev 113:109244. https://doi.org/10.1016/j.rser.2019.109244

Lopez-Castroman J, Moulahi B, Azé J, Bringay S, Deninotti J, Guillaume S, Baca-Garcia E (2019) Mining social networks to improve suicide prevention: a scoping review. J Neurosci Res. https://doi.org/10.1002/jnr.24404

Meng JCS (2009) Donald schön, herbert simon and the sciences of the artificial. Des Stud 30(1):60–68

Meziane F, Rezgui Y (2004) A document management methodology based on similarity contents. Inf Sci 158:15–36. https://doi.org/10.1016/j.ins.2003.08.009

Ngai EW, Lee PTY (2016) A review of the literature on applications of text mining in policy making. In: PACIS, p 343

Palm J, Thollander P (2010) An interdisciplinary perspective on industrial energy efficiency. Appl Energy 87(10):3255–3261. https://doi.org/10.1016/j.apenergy.2010.04.019

Petri I, Rezgui Y (2020) Bim for energy efficiency-decarbonising the built environment through informed decision-making using digital simulation and analysis

Petri I, Beach T, Rezgui Y, Wilson IE, Li H (2014) Engaging construction stakeholders with sustainability through a knowledge harvesting platform. Comput Ind 65(3):449–469. https://doi.org/10.1016/j.compind.2014.01.008

Petri I, Kubicki S, Rezgui Y, Guerriero A, Li H (2017) Optimizing energy efficiency in operating built environment assets through building information modeling: a case study. Energies 10(8):1167. https://doi.org/10.3390/en10081167

Petri I, Alhamami A, Rezgui Y, Kubicki S (2018) A virtual collaborative platform to support building information modeling implementation for energy efficiency. In: Working Conference on Virtual Enterprises, Springer, pp 539–550, https://doi.org/10.1007/978-3-319-99127-6

Rezgui Y (2007) Text-based domain ontology building using tf-idf and metric clusters techniques. Knowl Eng Rev 22(4):379–403. https://doi.org/10.1017/S0269888907001130

Rezgui Y, Miles J (2010) Exploring the potential of sme alliances in the construction sector. J Constr Eng Manag 136(5):558–567

Rezgui Y, Miles J (2011) Harvesting and managing knowledge in construction: from theoretical foundations to business applications. Routledge. https://doi.org/10.4324/9780203876091

Robinson G (2013) Global construction 2030: a global forecast for the construction industry over the next decade to 2030, vol 3. Oxford Economics

Sakaki T, Okazaki M, Matsuo Y (2012) Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans Knowl Data Eng 25(4):919–931. https://doi.org/10.1109/TKDE.2012.29

Singh T (2016) A comprehensive review of text mining. Int J Comput Sci Inf Technol 7(1):167–169

Singh N, Hu C, Roehl WS (2007) Text mining a decade of progress in hospitality human resource management research: Identifying emerging thematic development. Int J Hosp Manag 26(1):131–147. https://doi.org/10.1016/j.ijhm.2005.10.002

Song J, Song TM, Lee JR (2018) Stay alert: forecasting the risks of sexting in Korea using social big data. Comput Hum Behav 81:294–302. https://doi.org/10.1016/j.chb.2017.12.035

Succar B, Sher W (2014) A competency knowledge-base for bim learning. Aust J Constr Econ Build Conf Ser 2:1–10. https://doi.org/10.5130/ajceb-cs.v2i2.3883

Wu W, Issa RR (2014) Key issues in workforce planning and adaptation strategies for bim implementation in construction industry. In: Construction research congress 2014: construction in a global network, pp 847–856, https://doi.org/10.1061/9780784413517.087

Zhang L, Ashuri B (2018) Bim log mining: discovering social networks. Autom Constr 91:31–43. https://doi.org/10.1016/j.autcon.2018.03.009