Heliyon 8 (2022) e12088

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

Overview and analysis of the text mining applications in the construction industry

Hang Yan [a], Mingxue Ma [b], Ying Wu [c], Hongqin Fan [d], Chao Dong [a,*]

[a] School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan, China
[b] School of Engineering, Design and Built Environment, Western Sydney University, Sydney, Australia
[c] School of Management Science and Real Estate, Chongqing University, Chongqing, China
[d] Department of Building and Real Estate, The Hong Kong Polytechnic University, Hong Kong SAR, China

ARTICLE INFO

Keywords:
Text mining
Construction industry
Systematic review
VOSviewer
Text mining applications

ABSTRACT

The data generation in the construction industry has increased dramatically. The major portion of the data in the architecture, engineering and construction (AEC) domain are unstructured textual documents. Text mining (TM) has been introduced to the construction industry to extract underlying knowledge from unstructured data. However, few articles have comprehensively reviewed applications of TM in the AEC domain. Thus, this study adopts a qualitative-quantitative method to conduct a state-of-the-art survey on the articles related to applications of TM in the construction industry which published between the year of 2000 and 2021. VOSviewer software was applied to provide an overview of TM applications regarding to the publication trend, active countries and regions, productive authors, and co-occurrence of keywords perspectives. Eight prime application fields of TM were discussed and analyzed in detail. Five key challenges and three future directions have been proposed. This review can help the research community to grasp the state-of-the-art of TM applications in the construction industry and identify the directions of further research.

1. Introduction

Recently, the construction industry is experiencing an unparalleled growth in its strengths of data generation and data storage [1, 2]. Along with the advances of data collection technology, an immense amount of data in the construction industry is collected from the use of sensor systems, radio frequency tags, and construction extranet [3]. However, a considerable portion of the data in AEC domain are available as unstructured data sources, such as text documents, digital images, and videos. Specifically, in the construction industry, text documents consist of contracts, change orders, design reports, field reports, accident records, and others are very popular and crucial for stakeholders [3, 4]. Enabling these valuable data to be collected, stored, processed, retrieved, can assist in planning, controlling, and decision making in a project. Consequently, knowledge extracted from the diverse types of textual data can facilitate positive decision-making and better performance of construction projects [5].

With an increase in the availability of textual data, how to efficiently extract useful and crucial knowledge from these textual data becomes a primary challenge. Traditionally, manual management of a large amount of textual data is tedious and laborious, which could be solved by text mining (TM) techniques to some extent. TM is defined as a set of techniques which can be used to explore unstructured data sources and discover potentially valuable schemas, models, trends, or laws from textual data sources [6]. TM methods have been successfully applied in different fields such as finance [7], education [8], and medical health [9].

Though TM techniques have been introduced to AEC domain since early 2000 [10], the development of which was still lagging behind other industries [11]. Although an increasing number of researches about TM applications in the construction industry have been published in recent years, it was found to be difficult to transfer the findings of existing studies to the practices [12]. Additionally, only a few researches have reviewed and explored TM applications in the AEC domains. Soibelman et al. [3] reviewed previous works about data management and analysis of unstructured data in construction management, such as textual documents, site pictures, and web pages. Yan et al. [13] presented a comprehensive literature review on applications of data mining techniques in AEC domain. However, these two studies only reviewed a few articles related to TM applications in the construction industry and fail to give a comprehensive literature review. Additionally, Baek et al. [14]

* Corresponding author.
E-mail addresses: chaodong@whut.edu.cn, memorize.cool@163.com (C. Dong).

https://doi.org/10.1016/j.heliyon.2022.e12088
Received 21 March 2022; Received in revised form 27 September 2022; Accepted 25 November 2022

conducted a review of text analytics in the construction industry which focused on the data source and analysis methods, whilst the discussion of application areas was limited. In this case, this research attempts to conduct a comprehensive literature review of TM applications in AEC domain by combining the quantitative and qualitative methods. VOSviewer tool is adopted to analyze and visualize the state-of-the-art efforts of existing research. Then, a qualitative analysis on the eight application fields discussed in depth. In line with this, current research gaps and promising research orientations about TM applications in AEC domains are identified.

2. Text mining

TM, a branch of data mining, is generally defined as a knowledge-intensive process in which a user analyzes a collection of documents over a period of time by applying a series of techniques [6]. The difference between TM and data mining is that TM techniques extract the knowledge from text documents rather than from formalized database records [6]. TM can extract potentially useful, interesting, or meaningful patterns from non-structured or semi-structured data [15]. For the semi-structured data like web pages, some makeup are internal and describes the document structure; some is external and gives explicit hypertext links between documents. These information sources give additional leverage for mining web documents. Thus, Web mining, a new direction in TM filed, adopts wrapper induction, document clustering, and determining the authority of Web document to improve the effectiveness of TM in semi-structured data by utilizing the extra information available in web pages.

The TM process usually contains two procedures: text preprocessing and knowledge extraction process.

2.1. Text preprocessing

In this step, the representative characteristics are identified and extracted from document collections so that unstructured data stored in documents collections can be transformed into a more explicitly structured form [6]. The basic processing operations involve tokenization, stop words removal, stemming, etc. After text preprocessing, a document-term matrix is established based on the weighting of each term. There are several ways to determine term weight, such as binary term, term frequency, and term frequency-inverse document frequency [6]. Due to the sparse and high dimensional characteristics of the document-term vector, feature selection or dimension reduction techniques such as latent semantic indexing (LSI), probabilistic latent semantic analysis (PLSA), and linear discriminant analysis (LDA) are also adopted before data analysis [15].

2.2. Knowledge extraction process

For TM applications, knowledge extraction usually includes four common tasks: text classification, text clustering, information extraction, and information retrieval. For the different levels of representation, the analysis of textual data is diverse. In the applications of text classification, text clustering, and information retrieval, text data are usually treated as a bag-of-word without the semantic information. In applications of information extraction, it would be preferable to represent textual data semantically so that more insightful analysis and mining can be performed.

● Text classification

Text classification is to classify a collection of text documents into some predefined categories (such as topics and subjects). The documents should have been primarily converted into a manageable representation (feature vectors) during text preprocessing step. Then, some machine learning algorithms, such as support vector machines (SVM), neural networks (NN), and decision tree classifiers are adopted to construct a classifier by discovering the attributes of categories from a group of pre-classified examples [16]. Finally, the created classifier could place a new document in the appropriate categories by using the created classifier [6].

●Text clustering

Text clustering is an unsupervised process which assigns each document in a collection to one or more groups [17]. Text clustering depends on the maximization of the intra-cluster similarity and minimization of the inter-cluster similarity. Specially, documents in the same cluster are highly similar to each other, but have a low similarity to documents in other clusters. K-means, hierarchical clustering, and EM algorithm are widely used methods for text clustering [6]. Unlike classification, the labels of each cluster are unknown, and their meaning is implicit in the groups. Therefore, one important task in text clustering is to give a meaningful label to each cluster by domain experts.

● Information extraction

Information extraction can be viewed as a limited format of natural language comprehension in which one can define a prior type of semantic information which would be extracted from the document collections. The primary goal of information extraction is to extract data from documents to fill in slots in a pre-defined pattern or spreadsheet [16]. Then documents can be described as a collection of entities and frames which could depict the relationships between the entities in a formal way. The information extraction task can be decomposed into a sequence of steps, including named entity recognition, shallow parsing, building relations, inferencing, coreference resolution, template filling, and merging [6]. Hidden Markov Models, conditional random fields, and long short-term memory (LSTM) are widely used in information extraction tasks (Hotho et al., 2005).

● Information retrieval

Information retrieval is a task, responding to a query or some keywords and then looking for relevant information from a large set of documents [18]. For comparison, a search engine query can be viewed as a brief document. Then, a query can be also transformed into a vector of values that can be compared to the measurements in databases documents. The query will be matched against all documents stored in the databases, and the most similar set of documents will be retrieved. The basic technique of information retrieval is similarity measurement: two documents are compared to estimate their similarity. Three types of similarity measurement methods are frequently used include the cosine similarity, word count and bonus, and shared word count [16].

3. Research method

This study applies a holistic method to review the existing researches on the TM applications in AEC domain. The first step is a bibliometric review of journal articles. The second step is to build the science mapping of the literature samples, followed by the qualitative analysis of eight TM application areas in AEC domain. The workflow of this study is demonstrated in Figure 1.

Bibliometric search incorporates three sub-steps to explore literature which fall within the scope of TM applications in AEC domain. Science mapping, aided by VOSviewer, was adopted to analyze crucial findings on the basis of the finalized literature samples. Finally, this study conducted an in-depth qualitative analysis on TM applications from eight aspects.

3.1. Bibliometric search

In September 2021, a bibliometric search was performed, based on the data collected from five databases including Web of Science, Science

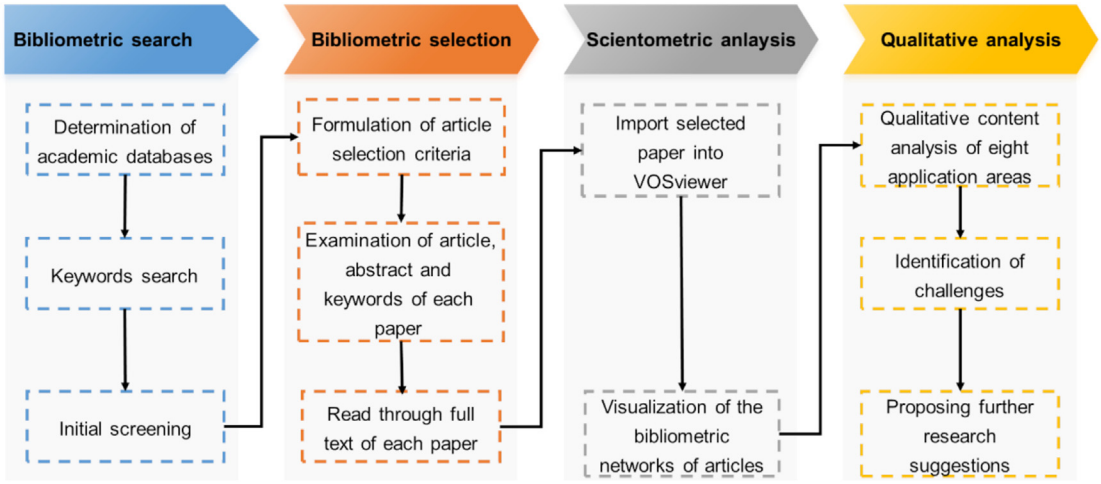

Figure 1. The workflow of research method.

Direct, Google Scholar, ASCE Library, and Engineering Village, which cover a great collection of high-quality journals and publications around the world. To identify studies related to TM applications in AEC domains, the keywords were as follows: TOPIC ("text mining" OR "textual data mining" OR "text analysis" OR "text classification" OR "text categorization" OR "text clustering" OR "information extraction" OR "text visualization" OR "information retrieval") AND TOPIC ("construction industry" OR "construction management" OR "AEC industry" OR "building" OR "built environment"). Only journal papers written in English were included in the search. A total of 1345 journal papers published between the years 2000 and 2021 were identified. To eliminate the possibility of duplicates, a second round of screening was carried out. Following the completion of the second round of screening, there were 437 articles left.

3.2. Bibliometric selection

Bibliometric selection includes two steps, namely, the formulation of article selection criteria and filter of articles. Studies should satisfy the following two criteria: (1) The primary research approach in this study should be TM techniques; (2) the study should be associated with TM applications in the construction industry. According to these two criteria, a two-phase selection strategy was implemented. In phase 1, a visual examination on important elements such as title, abstract, and keywords of each paper was carried out to extract the literatures that meet two criteria. In phase 2, another round of visual examination on the full text of studies was employed, to remove any studies whose contents show non-relevance to research topic. Finally, 127 articles from 38 journals were selected for further analysis.

3.3. Scientometric analysis

VOSviewer, a software for creating and visualizing bibliometric networks of articles, was adopted for scientometric analysis. Additionally, it provides TM capabilities for building and visualizing co-occurrence networks of key terms extracted from a collection of scientific publications [19]. Currently, the application of VOSviewer in scientometric review in the field of construction and project management is not special [20, 21, 22].

In this study, VOSviewer was applied to: (1) import the selected papers; and (2) visualize and compute the impacts of publication trends, active countries/regions, and productive authors. Seven types of scientometric analysis were conducted to present an overview of the literature sample, including the publication trend, active countries/regions, journal distribution, productive authors, co-authorship analysis, co-occurrence of keywords, TM methods, and TM tools. The content of these seven categories offers a comprehensive overview of the existing researches in related fields and usually reviewed by other researches [23, 24, 25].

3.4. Qualitative analysis

Following the bibliometric analysis and the science mapping, an in-depth qualitative analysis was carried out to analyze the TM applications in the AEC domain. Then, this qualitative analysis could identify the challenges in applying TM in AEC domain and provide suggestions for further study.

4. Overview of TM in construction industry

4.1. Publication trend

Annual publication trend on TM applications in the construction industry during the period from 2000 to 2021 is shown in Figure 2. It should be mentioned that this figure only describes the publication years of the 127 identified articles. As depicted in Figure 2, the last two decades could be further separated into three distinct periods: 1) 2000 to 2012, the publications related to TM in the construction industry is relatively few, with yearly number of publications not exceeding 4. This indicates that TM techniques were not widely applied in AEC domain during this period; 2) 2014 to 2018, the number of publications hovered between 5 and 11, which indicates a growing interest in the TM technique research in the construction industry; 3) 2019–2021, the research interest in TM applications have increased dramatically in AEC domain. With the maturation of TM approaches over the past two decades, an increasing number of scholars have undertaken to investigate TM applications in the construction industry.

4.2. Countries and regions

The researches on TM applications in the construction industry could be under worldwide cooperation. It could result in further analysis on active countries in related fields and specific impacts of this topic on the background of certain countries. Figure 3 describes the most active countries in investigating TM applications in AEC domain. According to Figure 3, a total of 19 countries/regions make contributions to TM research in the construction industry. The top 5 productive countries/regions were the United States (51 publications), China (36 publications), South Korea (13publications), the United Kingdom (8 publications), and Canada (6 publications).

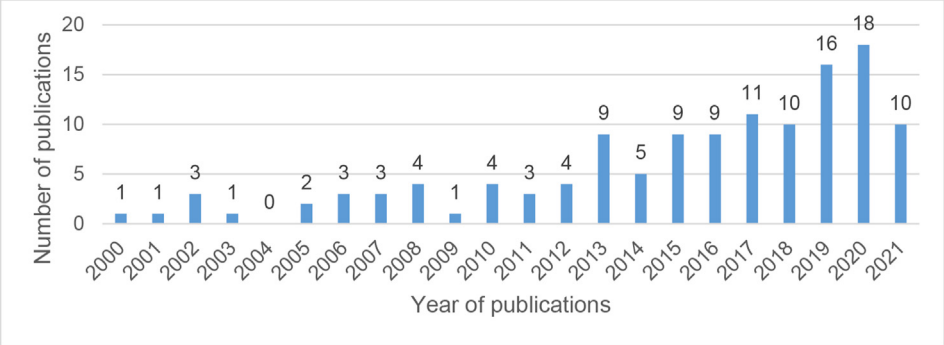

Figure 2. Yearly publications from 2000 to 2021. *Note: the number of journal papers in 2021 is incomplete because the articles selected in 2021 were only up to the September of 2021.

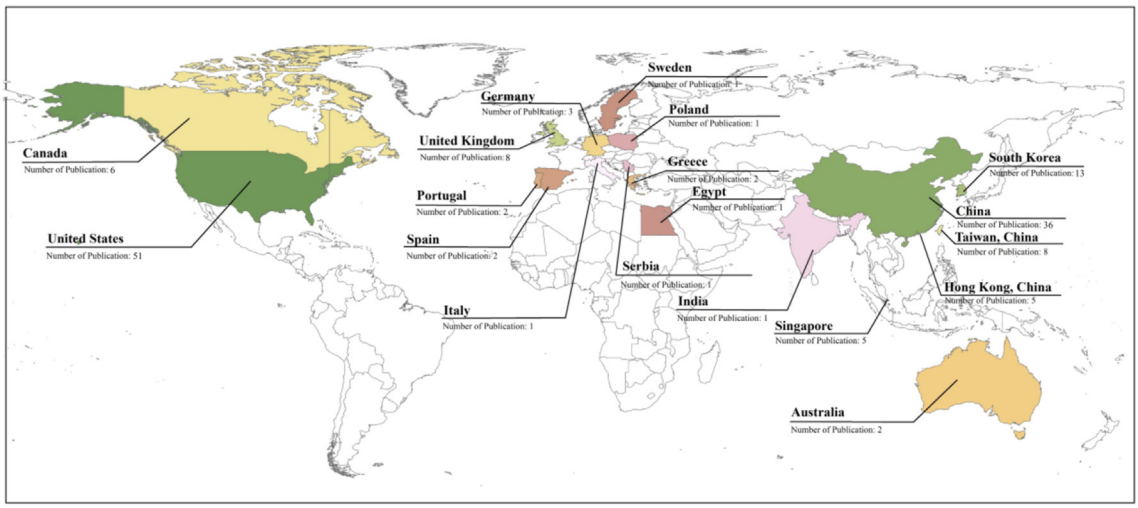

Figure 3. Publications distributed by country.

4.3. Distribution by journal

The total reviewed 127 publications are distributed in 38 journals. Among these journals, Automation in Construction (AIC, 29) published the greatest number of articles related to TM applications in AEC industry. The following are Journal of Computing in Civil Engineering (JCCE, 28), Journal of Construction Engineering and Management (JCEM, 11), Advanced Engineering Informatics (AEI, 9), and Journal of Management in Engineering (JME, 4). These top six journals focus on different research areas. AIC and JCCE are famous and influential in the field of civil engineering and information technology. The articles published by JCEM and JME are related to construction management. Additionally, AEI is a prominent journal in the research area of engineering informatics. The top 15 journals that have at least two articles are shown in Table 1.

Table 1. Distribution of the articles.

Rank	Journal	Number of articles
1	Automation in Construction	29
2	Journal of Computing in Civil Engineering	28
3	Journal of Construction Engineering and Management	11
4	Advanced Engineering Informatics	9
5	Journal of Management in Engineering	4
6	Expert Systems with Applications	3
7	Journal of Legal Affairs and Dispute Resolution in Engineering and Construction	3
8	Accident Analysis and Prevention	2
9	Building and Environment	2
10	Building Research & Information	2
11	Computing in Civil Engineering	2
12	Construction Management and Economics	2
13	Engineering, Construction and Architectural management	2
14	Journal of Architectural Engineering	2
15	Journal of Civil Engineering	2

4.4. Authors

The number of authors contributed to TM applications in AEC industry is 322. Table 2 demonstrates the key researchers who have contributed to the majority of review articles. The top 10 productive authors include: Nora El-Gohary (12 publications), Lucio Soibelman (7 publications), Ken-Yu Lin (6 publications), Botao Zhong (5 publications), Amr Kandil (5 publications), Jiansong Zhang (5 publications), Hanbin Luo (4 publications), Limao Zhang (4 publications), Heng Li (3 publications), and Weili Fang (3 publications).

To further reveal the relationship between authors and influential research teams, VOSviewer is used to conduct the co-authorship analysis. The minimum number of publications per author was set to 1 in performing co-authorship analysis. The co-authorship network between authors is displayed in Figures 4 and 5.

According to Figure 4 and Figure 5, it can be found that some groups of authors are more active and collaborative in past 20 years, such as the group of Heng Li, Zezhou Wu, and Ying Wang, the cluster of Botao Zhong,

Table 2. The top 10 productive authors from 2000 to 2021.

Rank	Author	Publications	Research directions
1	Nora El-Gohary	12	Ontology, NLP, deep learning techniques, construction management
2	Lucio Soibelman	7	Knowledge management, NLP, Computing in civil engineering, Machine learning, artificial intelligence (AI), and computer vision
3	Ken-Yu Lin	6	Construction health and safety, information technology, knowledge management
4	Botao Zhong	5	Construction safety, deep learning, blockchain, computer vision, text mining
5	Amr Kandil	5	Construction management, NLP, wastewater, text mining
6	Jiansong Zhang	5	Building information modeling (BIM), knowledge modeling, AI, NLP
7	Hanbin Luo	4	Deep learning, construction safety, digital twin, infrastructure construction, BIM, computer vision
8	Limao Zhang	4	BIM data analytics, AI, construction safety, risk analysis
9	Heng Li	3	Construction informatics, construction health and safety, BIM, construction management
10	Weili Fang	3	Construction safety, engineering informatic, infrastructure, deep learning, computer vision

Hanbin Luo, Xuejiao Xing, Junqing Tang, and Qirui Zhou, the group of Hongqin Fan, Hang Yan, Ya Wu and Fan Xue, the cluster of Fang Weili, Xing Pan, Lieyun Ding, and Jiexun Shuang, the cluster of Jintao Lin, Yi Wei, Lei Cao, Kun Lei, and Zhiling Yang, as well as the collaboration among Azizan Azia, Yunjeong Mo, and Matt Syal. It can also be found that each network group has its own critical focus area. For instance, the research group including Fan Hongqin, Liyin Shen, Wu Ya, and Xue Fan focused on applying TM techniques to retrieve information from similar cases for solving dispute and supporting green building design [26, 27].

4.5. Co-occurrence of keywords

Keywords exhibit the primary contents of articles and demonstrate the hot topics within a specific research area [[28]]. A keywords network presents the knowledge of these research topics' relationships, patterns, and intellectual organization [[29]]. Among the 127 identified articles, "Keywords" and "Full Counting" are presented in the VOSviewer. In the original output, the occurrence number of 35 out of 395 keywords are more than 5. Two criteria that filtering keywords are proposed: (1) the general keywords such as "internet", and "documentation" are removed; and (2) semantically consistent keywords are merged into one term, for example, "TM" and "text-mining". Finally, a total of 21 keywords are identified. Their visualization is presented in Figure 6 and Figure 7.

In Figure 6 and Figure 7, the font and node size indicates the occurrence number of each keyword in the reviewed articles. The edge between nodes denote their inter-relatedness. According to Figure 6, some keywords indicate the prevailing techniques related to TM, including data mining, ontology, NLP, information management, and machine learning. In addition, the mainstream application areas in AEC domain are also identified, such as construction safety, building information modeling (BIM), automated compliance checking, and topic modelling. These keywords are categorized into several groups and connected to one another via connection lines. For instance, keyword of "text classification" often coexists with "machine learning" and "NLP", which means researchers usually adopt these two techniques to perform text classification.

Figure 7 presents the evolution of main research topics about the TM applications in AEC domain. The brighter color of the node means the more recent publication year of this topic. It can be found that "machine

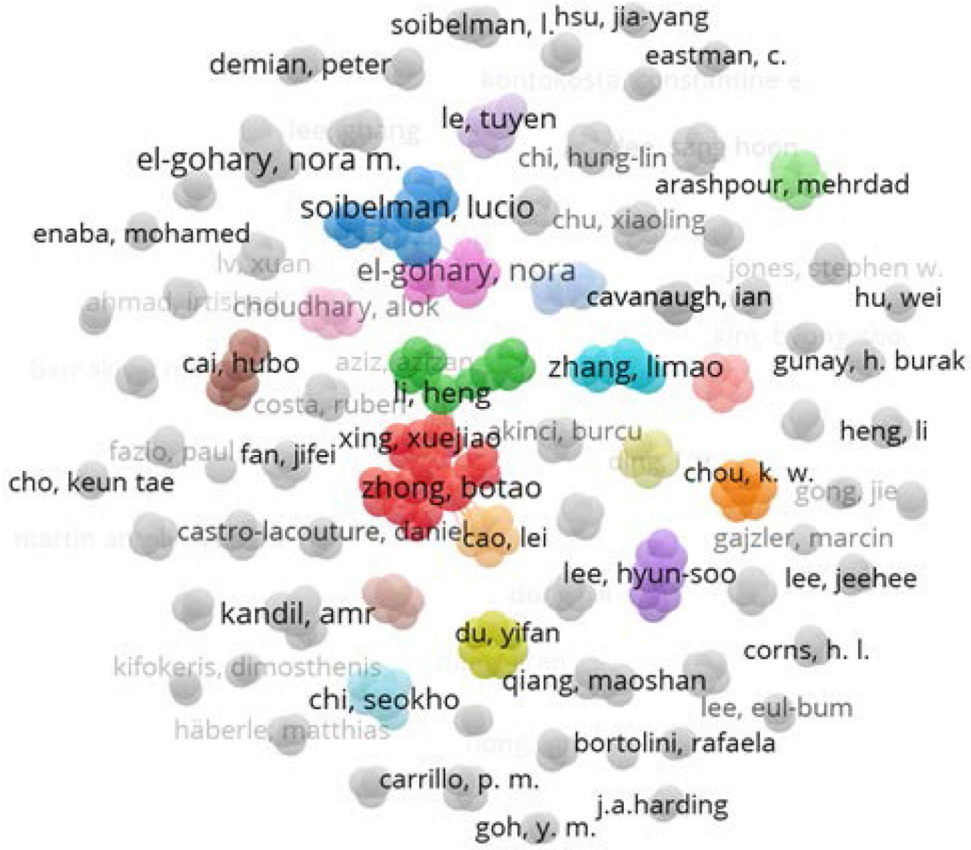

Figure 4. Co-authorship analysis of authors regarding to TM in the construction industry from 2000 to 2021 showed a total of 322 nodes, and 70 clusters.

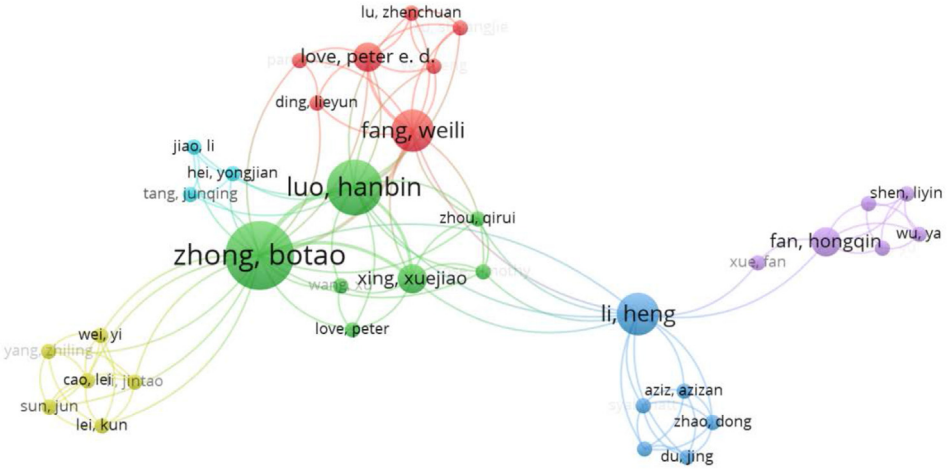

Figure 5. The largest set of connected items in co-authorship analysis of authors regarding to TM in the construction industry from 2000 to 2021 consists of 35 nodes and 6 clusters.

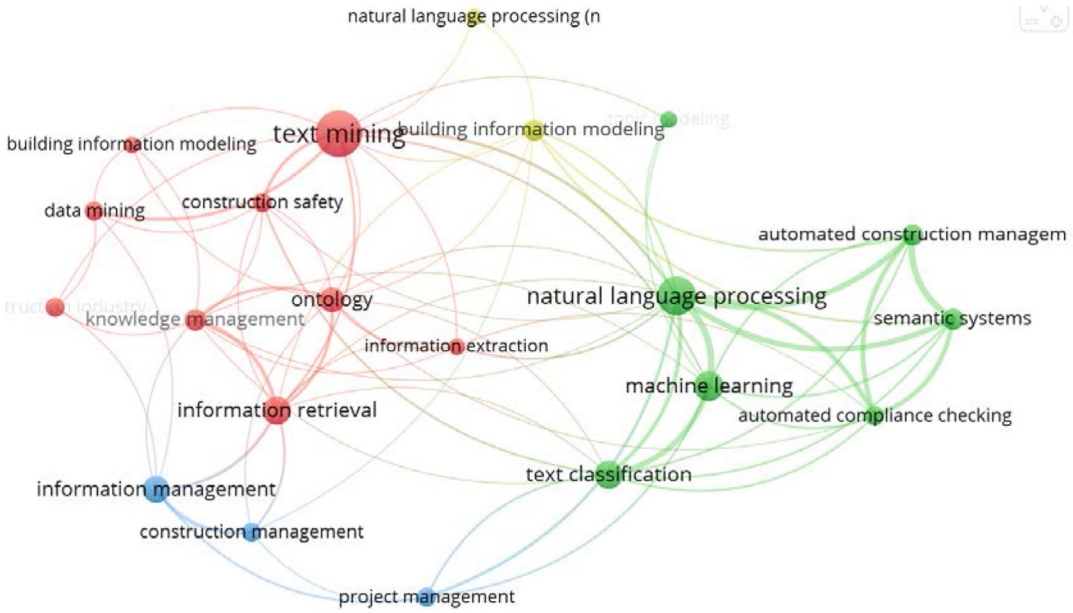

Figure 6. Co-occurrence of keywords analysis regarding to TM in the construction industry from 2000 to 2021 showed a total of 21 nodes, and 4 clusters.

learning", "topic modelling", "NLP", "information extraction", and "construction safety" are the latest research topics, while topics such as "information management", "construction management", and "information retrieval" are relatively outdated in this field.

4.6. TM methods

The TM methods or algorithms used in the reviewed literature are presented in Table 3. For information retrieval, Boolean model and the vector model are the relatively common methods. For information extraction, the typical methods include Latent Dirichlet Allocation (LDA) and rule-based method. For text classification, Support Vector Machine (SVM), Navie Bayes (NB), K-Nearest Neighbor (kNN), DT, and logistic regression are the most common methods. Furthermore, the performance of these classification methods is usually compared to discover the most appropriate methods for a specific task. For example, Qady and Kandil [30] compared the classification performance of Rocchio classifier, NB, KNN, SVM and concluded that Rocchio classifier and KNN achieve higher accuracy. Hussan and Le [16] applied NB, SVM, logistic regression, and feedforward neural network to classify the contractual requirements. This research finds that SVM have performed best in the aspects of accuracy, precision, recall, and F1-score. For text clustering task, K-means are adopted by the most researchers.

4.7. Tools used for TM applications

As shown in Table 4, various tools have been adopted by TM applications in AEC industry, including customized software, high-level languages, TM software, statistical software, and special purpose software.

Among all the tools, high-level languages are the most commonly used in the reviewed literature, especially Python. Python is an interpreted high-level general-purpose programming language. It is concise, powerful, readable, friendly, and easy to learn. Writing programs in Python takes less time than in other common languages like C, C++, and Java. Moreover, large amounts of third-party toolkits such as, NLTK, BeautifulSoup, Pandas, TensorFlow, can be used directly to implement different TM functions. Therefore, Python is popular in TM field.

Well known TM software including STATISTICA, IBM Miner for Text, Rapid Miner, PolyAnalyst, can offer some core functions of TM such as text classification, text preprocessing, and information extraction.

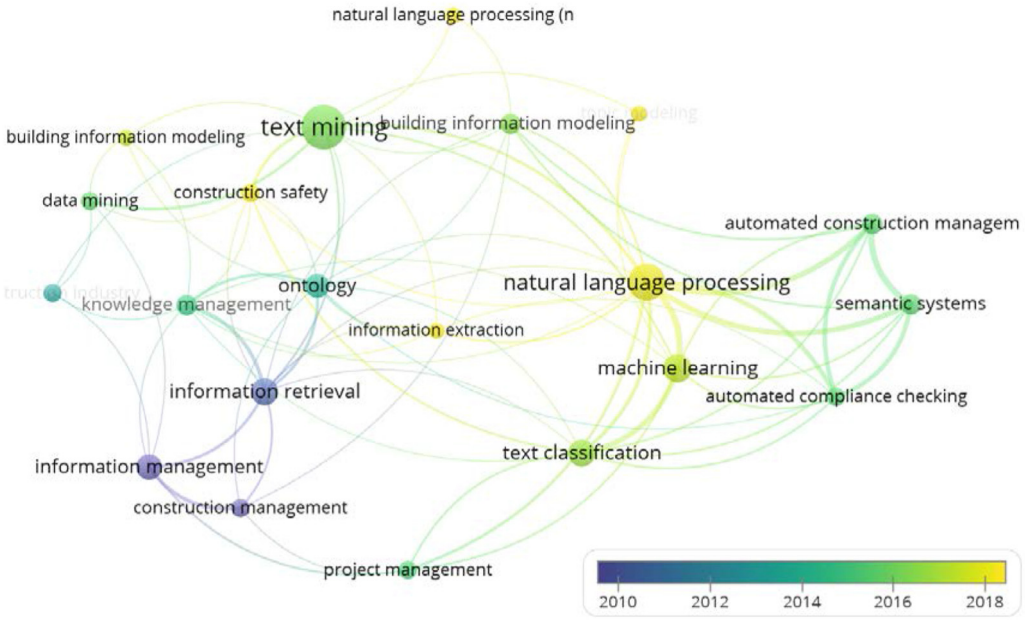

Figure 7. The evolution of main research topics regarding TM in the construction industry from 2000 to 2021.

However, these tools are less flexible than the high-level language since they cannot customize the functions for users. Additionally, some statistical software also has some functions of TM such as WeKa, SPSS, Matlab, and SAS. These tools have the similar shortages with the TM software.

Special purpose software like protégé, WordNet are commonly used for ontology construction and reasoning [31, 32, 33]. Additionally, some scholars developed customized software to achieve some special functions which cannot be accomplished through existing commercial or open source software. For example, Yeuny et al. [34] developed a narrative knowledge representation system (NKERS) to fulfill the extraction and representation of knowledge in the construction industry.

5. Application areas of TM

TM has been widely applied in diversity aspects in AEC domain. This study identifies eight application fields where TM techniques are popularly used. The application fields include safety management, automation compliance checking, public opinion analysis, building design, framework development, method development or improvement, contract management, and others. A deep qualitative analysis is conducted for these eight application fields. Four TM functions which are widely used in these application fields for different purposes are also identified (See Table 5).

5.1. Safety management

Safety is a critical topic in the construction industry. In all reviewed articles, TM techniques are most frequently used in safety management areas. Specifically, TM techniques can retrieve similar accident cases, identify factors and causes of accidents, and classify accident reports [35].

Several researches have conducted information retrieval from the accident reports. For example, Fan and li [27] applied TM techniques to retrieve historical cases of dispute resolution from a case base. Based on this fundamental research, the retrieve system proposed by Kim et al. [36] can extract BIM objects and create a query set by merging BIM objects with a project management information system (PMIS). Zou et al. [37] attempted to expand user's query based on pre-defined risk vocabulary and WordNet in their risk case retrieval system. Kim and Seokho [38] not only expanded query by utilizing a lexicon of construction accident cases, but also extracted tacit knowledge from accident cases by applying rule-based and conditional random field (CRF) methods.

For classifying accident reports, supervised learning and unsupervised learning methods were both used in previous studies. For example, Kifokeris and Xenidis [39] applied an unsupervised clustering algorithm on a vast risk notion set to deduce risk sources of a project. Chi et al. [40] adopted SVM to classify job hazard under different predefined safety violation scenarios. Cheng et al. [41] introduced a supervised machine learning method named Symbiotic Gated Recurrent Unit (SGRU) to classify site accident reports. Some scholars also evaluated the performance of various machine learning algorithms for classifying accident reports [42, 43, 44]. Furthermore, the application of deep learning methods become popular and trendy in classifying accident reports [45, 46]. Additionally, ontology techniques were integrated in job hazard analysis to improve the efficiency of text classification [47].

For information extraction tasks, Martínez-Rojas et al. [48] proposed a rule-based methodology for automatically extracting some requirements from safety and health plans. Tixier et al. [49] also designed a rule-based model to extract structured features from unstructured accident documents. Choi and Cho [50] adopted TM and network analysis to extract the keywords highlighted by CEOs and the key factors of safety management. Baker et al. [51] compared different approaches to automatically extract injury precursors from project accident documents.

5.2. Automation compliance checking (ACC)

Construction projects are in the control of numerous regulations including norms, laws, regulations, corporate policies, and contract terms [52]. Ensuring compliance of their work against the norms is one major task to experts in the construction industry [31]. However, inherent complexities of these regulations are one critical obstacle in a successful compliance checking due to the variability of their provisions and depth and breadth of domain knowledge [53]. The compliance checking based on manual approach is criticized as a tedious, expensive, and error prone process [54]. Therefore, ACC, as a more efficient approach, is anticipated saving time, money, and eliminate errors.

Generally, a typically ACC process consists of four basic steps (as shown in Figure 8): (1) extracting rules from regulations, contracts, and advisory reports and standardizing these rules into logical phrases; (2)

Table 3. Number of publications by TM methods.

Function	Method	Number
Information retrieval	BM25 method and thesaurus weight	1
	Similarity measure	2
	Boolean model	4
	Vector model	6
Information extraction	Latent Dirichlet Allocation	7
	Rule-based method (semantic mapping rules, conflict resolution rules, rule based chunking, pattern-matching-based rules, lexicon-based method)	10
	shallow parsing (CRISP)	1
	Bidirectional long short-term memory (Bi-LSTM)	1
	Conditional random field (CRF)	1
	Link analysis	1
	Generalized Suffix Trees	1
	Probabilistic latent semantic analysis (PLSA)	1
	Topic-over-time (TOT) model	1
	Bidirectional encoder representations from transformers (BERT)-based sentiment model	1
	conditional random field (CRF)	1
	Doc2Vec	1
Text classification	MapReduce algorithm	1
	Deep Belief Network (DBN)	1
	Latent semantic analysis (LSA)	2
	Latent Dirichlet allocation (LDA)	1
	Support vector machine (SVM)	24
	Rocchio algorithm	3
	naive Bayes	18
	k-nearest neighbors	13
	Radius-based neighbors	1
	Nearest centroid	1
	Random forest	4
	Decision tree	9
	gradient boosted regression trees	1
	Feedforward neaural network	1
	Logistic regression	6
	Rule-based classification	1
	Bidirectional Transformers for Language Understanding (BERT)	1
	Convolutional Neural Network (CNN)	4
	Hierarchical Attention Network	1
	Long Short-Term Memory Neural Network (LSTM NN)	3
	Single word and Bi-gram models	1
	Projective adaptive resonance theory (PART)	1
	Frequent item analysis (FIA)	1
	Hierarchical softmax skip-gram algorithm	1
	Radior rules	1
	K-star	1
	Radial Basis Functions	1
	Stacking	1
	Gated Recurrent Unit (GRU)	1
	Symbiotic Gated Recurrent Unit (SGRU)	1
	Multiple Key Term Phrasal Knowledge Sequences (MKTPKS)	1
Text clustering	k-means	4
	Single pass clustering algorithm	1
	Efficient Fuzzy Kohonen Clustering Network (EFKCN)	1
	Ward.D method	2
	Expectation Maximization	1
	Spring Embedding	1
	Agglomerative hierarchical clustering	1
	Gaussian mixture model-based clustering	1
	fastText algorithm	1
	Partition around medians (PAM)	1
	hierarchical clustering	1

extracting and standardizing the information collected from a specific project; (3) conducting the compliance checking between the rules and project information; and (4) outputting the results of the compliance checking. The first two processes of extracting and standardizing rules involve two TM functions, namely, information retrieval, and information extraction [52].

With the development of text analysis technology, many research efforts have been spent in ACC area. These research efforts include: (1) using encoded rules for automatic checking of building designs [54,55, 56,57]; (2) exploring a semantic modeling method based on ontologies to facilitate compliance checking on construction quality [58]; (3) adopting a rule-based semantic method as a strategy for automating regulatory compliance [31, 59]. Nevertheless, these scientific explorations are restricted to their automation and reasoning capabilities in ACC process. A research team from University of Illinois has conducted a series of studies to solve these problems. This research team proposed a new approach for ACC task based on deontology, deontic logic, and NLP techniques [60]. In order to enhance the efficiency and accuracy of information extraction, they proposed multiple methods to classify regulation documents before ACC process, such as machine learning-based text classification method [61, 62], and an ontology-based text classification algorithm [63]. Then, a rule-base NLP approach was introduced to extract valuable information from construction regulatory files automatically [53]. Additionally, Zhang and EI-Gohary [64] proposed an innovative approach for the extending IFC schema to involve the information of compliance checking and support ACC.

5.3. Public opinion analysis

The popularity and prosperity of the Internet attracts more and more individuals to engage in social media. Real-time data from blog posts, status messages, posts, and comments have potentially provided valuable information for administrator [65]. However, data from social media is big, unstructured, heterogeneous, and with a lot of noise. It is very difficult to process, store, and analyze a huge amount of data from social media by using manual approach [66]. To address this issue, TM-based approaches have been introduced to extract potential social phenomena from social media data. A general flow chart for data analysis of social media is described in Figure 9.

Firstly, the keywords related to a specific topic should be determined to search the target posts on the social network. Then, the web crawler, a data collection technology, can automatically collect target data from one or more webpages based on a certain strategy [67]. Then, data preprocessing techniques (i.e. tokenization, stopping, stemming) are used to transform target data into a structured format.

Sentiment analysis, also called opinion mining, is a method to acquire netizens' emotions, viewpoints, and attitudes from social media [66, 68]. In addition, machine learning-based and lexicon-based methods are two popular types of technologies in the sentiment analysis. Specifically, machine learning method uses typical classification algorithms such as Naïve Bayes, SVM, and the Maximum Entropy to classify sentiment words and themes [69]. Lexicon-based approach combines a sentiment lexicon with a series of emotion words, for example, interest, happiness, anger, sadness, surprise, fear, disgust [70], with pre-established regulations to complete the sentiment analysis. Two basic subtasks of sentiment analysis are sentiment value calculation and sentiment orientation analysis [68, 71].

Table 4. Software used for accomplishing TM tasks.

Type of tools	Name of tools	Number	Used for
High-level language	Python (NLTK, Numpy, scikit-learn, lda, Selenium, BeautifulSoup, Pandas, TensorFlow, matplotlib, gensim, torch, Gensim, pyLDAvis, Plotly)	25	Information retrieval, data preprocessing, Topic modelling, information transformation, text classification, information extraction, similarity measure, text visualization
	Java (GATE, B-Prolog's, LJB, MALLET, LIBSVM, JSDAI, Solibri Model Checker)	13	System development, NLP, information extraction, compliance reasoning, data preprocessing, text classification,
	C++	4	System development, automated building code checking
	PERL	1	Search engine
	Microsoft SQL Server	1	Text preprocessing
TM software	STATISTICA	1	Data preprocessing
	IBM Miner for Text	1	Text classification
	General Architecture for Text Engineering (GATE)	3	Text classification, information extraction
	PolyAnalyst	2	Text preprocessing
	CLUTO	1	Text clustering
	KoNstanz Information MinEr (KNIME)	1	Text preprocessing, information extraction
	JDOM	1	Parser development
	Rapid Miner	2	Text classification, text preprocessing
	Stanford Nature Language Processing (StanfordNLP)	3	Information retrieval, sentiment analysis
	Institute of Computing Technology Chinese Lexical Analysis System (ICTCLAS)	2	Text preprocessing
	ROSTCM	1	Word Cloud Visual Analysis
	SAS Text Miner	1	Text clustering
	VOSviewer	1	Literature analysis
	Netdraw	2	Social network analysis
	Voyant-tool	1	Text preprocessing
	EXPRESS Data Manager (EDM)	1	Rule checking
Statistical software	SPSS	1	Association analysis
	R (TM, SnowballC, JiebaR, MCLUST)	5	Web crawling, Data Preprocessing, text clustering
	Gephi	1	Network analysis
	Octopus	2	Web crawler
	Natural language process & information retrieval (NLPIR) big data analysis platform	2	Sentiment analysis
	Weka	4	Text classification, text clustering
	Matlab (Text Analytics Toolbox)	4	Text pre-processing, sentiment analysis, text visualization
	nPlan software	1	Text clustering
Specific purpose software	Protege	4	Ontology construction, query, reasoning
	WordNet (Princeton University)	1	lexical database of the English language
Customized software	OntoPassages	1	Information retrieval
	A narrative knowledge extraction and representation system (NKERS)	1	Knowledge extraction and knowledge representation
	Construction document classification system (CDCS)	1	Text classification
	UNI-Tacit	1	Information extraction, text visualization

Table 5. Number of articles for each application field.

Application fields	Number of articles	TM functions	Number of articles
Safety management	21	Information retrieve	5
		Information extraction	6
		Text classification	8
		Text clustering	2
Automation compliance checking (ACC)	13	Information extraction	11
		Text classification	2
Public opinion analysis	11	Information extraction	11
Building design	11	Text classification	3
		Text clustering	1
		Information retrieval	3
		Information extraction	4
Framework development	11	Text classification	2
		Information retrieval	8
		Information extraction	1
Method development or improvement	25	Text classification	9
		Text clustering	3
		Information extraction	4
		Information retrieval	9
Contract Management	12	Text classification	1
		Information extraction	9
		Information retrieval	2
Others	23	Text classification	5
		Text clustering	2
		Information extraction	15
		Information retrieval	1

Topic modelling is another important task in the analysis of social media. LDA and PLSA are two famous algorithms to discover topics [72, 73]. Generally, topic modelling attempts to count terms associated with topics, in order to discover the appearance of abstract "topics" in a set of documents [74]. It assumes that a specific topic derived from textual document set can be represented by a group of keywords. In public opinion analysis, topic modelling is usually adopted to discover the key focus of public attention from a large collection of posts on a specific topic [68, 71].

In the construction industry, public opinion or owners' feedback have an influence on the construction and maintenance of a project. In planning and design stage, collecting and analyzing public-opinion information can provide valuable information for decision making, especially for

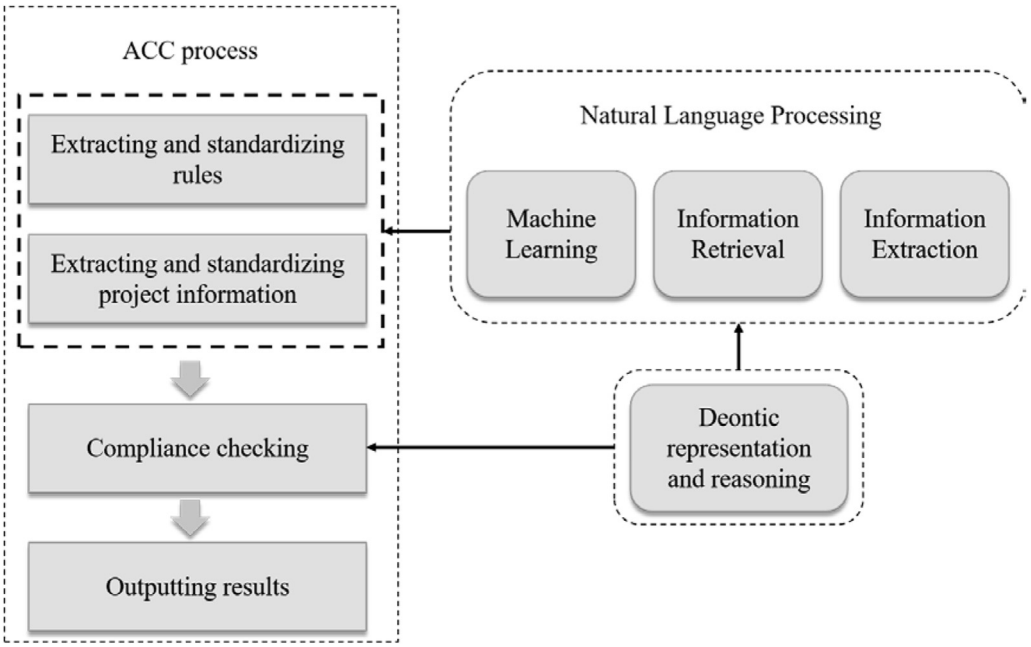

Figure 8. The typically processes in ACC.

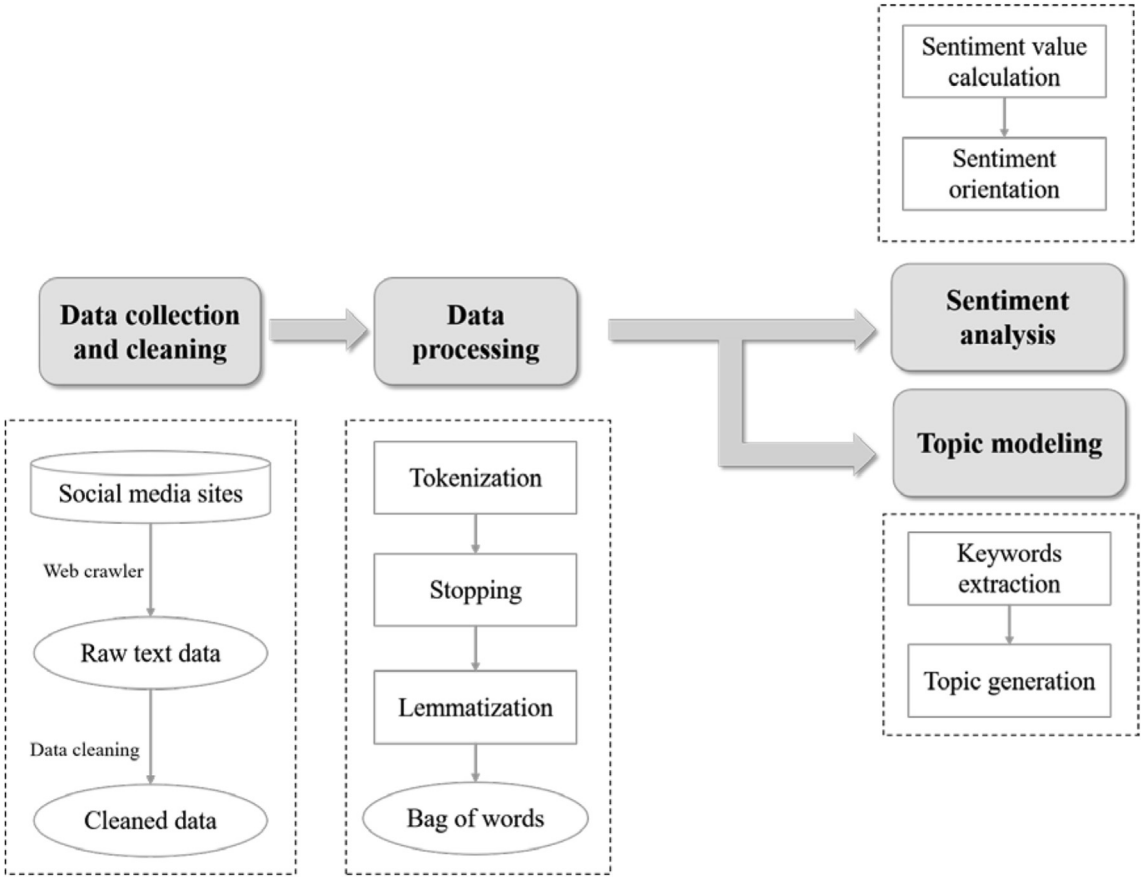

Figure 9. A general flow chart for data analysis of social media.

the infrastructure projects, which can benefit a large group of people. For infrastructure projects, the public are willing to actively engage in the decision-making and strategic planning steps [75]. In the operation and maintenance stage, a substantial amount of end-user maintenance feedback might be generated. Analyzing end users' feedback can investigate the common fault symptoms and problems in property management [76]. Furthermore, public opinion on some specific topics in AEC domain, including green building [71], off-site construction [68], or US construction industry [77], can help policy maker gain a better understanding of public attitudes and sentiment orientation. Additionally, TM

techniques could be used to capture users' needs on BIM applications and indoor environmental quality [12, 78].

5.4. Building design

Due to the non-deterministic and subjective nature of design behaviors, the content-based TM approach is a useful decision-making instrument for evaluating design productivity and creating individualized work arrangements for a more effective modeling process [79]. It also provides a promising solution for text classification, information extraction, text clustering, information retrieval of vast design documents.

From the data sources perspective, the focus of TM applications in building design is BIM event log data. BIM design logs can automatically save specific design information of the modeling step, which provide data sources for knowledge extraction [80]. Some researches classified BIM log data improve the design efficiency and rectify design errors by using various methods, such as LSA and LDA [81], LSTM NN [79], and EFKCN [79]. Some potential patterns, knowledge, features were extracted from BIM log data, such as social networks [82], sequential patterns [83], and features of BIM model [84]. Other data sources like CAD and tradition architect reports were also used in TM applications [85] and Zhang et al. [83].

A significant TM application in the building design is to retrieve similar design schemes for references when creating new design plans [26, 86]. This can help architects enhance the efficiency and effectiveness of building design.

5.5. Framework development

Framework development in AEC domain relies on the diversity of TM techniques and other related techniques, including web crawling, ontology, and case-based reasoning (CBR). These frameworks generally cover the whole process of knowledge discovery, including data gathering, data processing, and results presentation. They have a wild range of applications and can be extended to solve multiple specific problems. Typically, case studies are conducted to imply the appropriateness of a framework. Some commercial TM platforms (such as hanLP, Baidu NLP, aliyun, and IBM Watson Natural Language Understanding) provide some basic TM toolkits for users to develop the analysis framework in AEC domain. Users with little knowledge of TM algorithms can also easily use these toolkits with the guide of platforms to complete the TM tasks like multi-language segmentation, named entity recognition, text classification, text clustering, etc.

For information retrieval systems, Lee et al. [1] established a framework for sharing defect data across disparate data sources. This framework consists of four steps: developing a defect ontology, extracting work related information from BIM models, transforming BIM data into RDF format, and implementing SPARQL. Liu et al. [87] proposed a domain-specific information retrieval system which can automatically collect, merge, and offer the information required by them. Fan et al. [32] designed a framework to retrieve project-wide as-need information from construction files. To improve the efficiency of information retrieval system, a lexicon for the specific project and dependency grammar parsing was adopted. Scherer and Reul [10] developed a knowledge system to retrieve project knowledge from construction documents. Kovacevic et al. [88] carried out a question-and-answer system where construction practitioners can pose questions and easily find information rather than imputing a list of keywords. For text classification systems, Caldas and Soibelman [89] established a document classification system to enhance information organization and access in inter-organizational systems. In addition, Caldas and Soibelman [90] elaborated a text classification method and improved data organization and accessibility in an information system of construction management. Meanwhile, a prototype of a document classification system was carried out by this research team to facilitate quick deployment and scalability of the classification step [91]. TM techniques were also used for knowledge acquisition and knowledge representation in document management systems [4, 92, 93].

5.6. Method development or improvement

Some scholars have explored TM techniques on theoretical level, including applications of these TM methods, integration with other methods for improvement. Particularly, ontology methodology was widely used to improve the efficiency of information retrieval [94, 95], information extraction [33, 96], and document classification [97]. Similar to framework development, these improved methods can be extended to other domains.

For document classification, Sebastiani [98] comprehensively discussed of the main text categorization approaches. Mahfouz [99] focused on one of the text classification methods (SVM) and its application in the construction industry. Qady and Kandil [30] proposed a hybrid approach which initially used single pass clustering to generate core clusters and trained a textual classifier on these core clusters to categorize outlier files in a subsequent refining process. This research team also compared the performance of four classifiers for automatically classifying construction files under varying general conditions [100]. For knowledge representation and extraction, Qady and Kandil [5] used shallow parsing (CRISP) to identify concept relations from construction contracts which can be applied to improve the performance of text classification and information retrieval. Nedeljkovic and Kovačevic [2] proposed a method to establish a network of critical phrases from construction documents, for less efforts spent on extraction and visualization of valuable project facts. Zhu et al. [101] captured the concepts associated with construction components, contract files, and construction management in unstructured content and constructed a metadata model of request for information problems. Zhong et al. [102] proposed a hybrid deep neural network by combining Bi-LSTM and CRF to automatic extract the construction procedural constraints. Kim et al. [103] specified a generic method to automatically draw out obvious semantic document structure from a structural calculation file. Le and Jeong [104] presented a novel method which integrates several computational algorithms to extract concepts and their relationships from design handbooks and other technical documents. In addition, this study also proposed an algorithm to classify semantically related words into three distinct lexical categories. For information retrieval, Demian and Balatsoukas [105] discussed the granularity concept in information retrieval systems when evaluating correlation and visualizing retrieved results. Lin and Soibelman [106] proposed two approaches, namely, query extensions and conceptual indexing in information retrieval models. Additionally, several methods were proposed for data collection [107], text visualization [108, 109], and the whole process of information management [110].

5.7. Contract management

Construction contract is a legal document of a building project that indicates all the requirements and expectancies of the owners during the life cycle of a project. It is critical to precisely understand contract documents to ensure all requirements are captured and managed [17]. On the other side, extracting and identifying some special clauses such as exculpatory clauses, change order, and poisonous clauses are also very helpful before signing a contract [111, 112, 113]. However, the traditional manual practice of requirement identification is time-consuming, tedious, and error-prone [114]. Therefore, TM approaches have been widely employed to extract and identify contractual requirements.

The TM function in most reviewed researches of the contract management is information extraction. But before extracting useful information from the contract, there is a need to classify the contractual content into requirements and non-requirement [17]. Then, researchers have made significant attempts to build rule-based models to extract hidden knowledge for contracts [113, 115]. However, rule-based models frequently depend on pre-defined lexicons and require intensive feature

engineering. Zhong et al. [116] introduced a deep learning-based method by combining Bi-LSTM and CRF to extract constraints without complex handcrafted features engineering. Furthermore, some complex models have been established for achieving multiple functions. For example, Marzouk and Enaba [117] not only extracted the critical terms from contracts but also identify the obligations of each party and cluster the project correspondence. Sun et al. [108] optimized TF-IDF algorithm by considering the characteristics of engineering texts when extracting information from the contacts. Then, this study also visualized the key information extracted from contract reports by using tag cloud algorithm, keyword centrality analysis, and multidimensional scaling and clustering analysis.

5.8. Others

As different types of documents are generated during the life cycle of construction project, TM techniques have a wide range of applications in AEC domains. In addition to the above areas, TM techniques have also been used in defect analysis [118, 119], material management [120], workforce planning [121], facility management [87, 122, 123], complaints management [124], contract management [113, 117], field inspection [125], product search [126, 127], post-project reviews [128, 129, 130, 131], cost prediction [132], quality management [68], job advertisements [133, 134], legal decision support [135], building renovation [136], feasibility study [137], patent analysis [138], schedule delay [139], and incident duration prediction [140].

6. Challenges and future work

As discussed above, applications of TM can eliminate existing problems in AEC domain and contribute to the achievement of considerable progress. Nevertheless, there are five challenges in this research area. Additionally, the corresponding future directions are proposed.

6.1. Challenges

● Limitations in ACC

While the performance of knowledge representation and reasoning is important to integrate norms, current ACC systems in the construction industry rarely provide it. Furthermore, the methods proposed by recent researches were limited to the simple form of rules, which are unable to perform more complex compliance checking. Manually extracting rules from norms is time-consuming and laborious. Additionally, the existing ACC systems are featured as inflexibility that users restrict to add or modify rules in these systems. Therefore, ACC is a challenge not only for researchers in the construction industry but also for artificial intelligence (AI) experts. With the development of AI technology, deeper and broader NLP algorithms will be adopted to analyze complicated sentences and discern latent meaning, the above-mentioned problems will be solved in the future.

● Limitations in the analysis of public opinion

TM techniques have been used to explore user or public sentiment in AEC domain through collecting data from social media. However, there are still some unsolved problems. Firstly, many terms that are not included in the sentiment dictionaries do express emotions when used in a specific topic. Secondly, it is extremely hard for a machine to comprehend the complex sentences with ironic meaning [141]. Thirdly, opinions from specific population groups, such as the elderly, may be neglected since the typical user of social media is between the ages of 18 and 35 [66]. Fourthly, people usually share their viewpoints on social media anonymously and therefore the data collected from social media may be fake [142]. Therefore, advanced techniques are in need to solve these problems. For example, a domain sentiment dictionary which covers a list of words, concepts, and phrases in the AEC field can be constructed to improve the accuracy sentiment analysis.

● Less utilization of domain knowledge

Despite some efforts have been made to incorporate domain knowledge into the TM process, the significance has not been strongly emphasized. Specifically, only a small part (nearly 10%) of reviewed articles integrated ontology techniques into TM applications. Insufficient domain knowledge in TM analysis will result in the problems of semantic ambiguity, information loss, etc. These problems will further negatively affect the performance of information extraction and information retrieval.

● Ignoring of structural characteristics

In order to simplify data processing, most TM applications in the construction industry ignored the structural features of the textual content (e.g., font type, size, location, etc.). Usually, the textual documents are presented as a bag of words which permit the algorithms to concentrate on the semantic features. Nevertheless, this approach can cause a loss of valuable information. For example, the title of a journal paper can be easily identified, due to its special font and specific location rather than its semantic content alone. The visual layout of documents can provide some frequently overlooked but critical information.

● Publicly available datasets

Data sources and data volume have very significant effects on TM model performance in the construction industry. Some data sources are available for public such as social media, regulations, standards, and government open databases. However, most of data sources in the reviewed articles are project-related data which is unavailable for public due to its confidentiality.

There is a trend toward releasing and offering more information for the public around the world [143]. Governments around the world (e.g. U.S. Data.gov, U.K. Find Open Data, European Data Portal) and international institutions (such as World Bank, OECD, and International Monetary Fund) have provided open data for public [14]. Although these open data sources will benefit for the TM applications in the construction sector to a certain extent, publicly available datasets specific for the construction industry are still insufficient. Therefore, it is urgent to appeal to the construction administration departments, relevant international institutions, and researchers around the world to open more data to improve knowledge discover in AEC domain.

6.2. Future research directions

● Integration of TM with domain knowledge or ontology

As discussed above, the performance of TM applications is undesirable without domain knowledge. Therefore, some scholars attempted to apply ontology to the TM systems. Ontology as an ideal tool to model knowledge representation is widely used in information science [144]. It is defined as a framework for knowledge representation of a specific area. It uses a range of concepts, relations, and axioms that allows semantic reasoning [145]. Ontology can erase the limitations of semantic ambiguity and enhance the performance of information extraction, information retrieval, and text classification. Concretely, use of domain ontologies in text preprocessing can create a uniform lexical reference space and corresponding hierarchical relations for concepts. This is helpful in subsequent queries, presentation, and refinement operations. Furthermore, domain knowledge can be crafted into constraints in contracts, regulations, and clauses which enables more efficient and significant queries. Therefore, ontology is an important technology and its combination with TM could be further employed to solve problems in

AEC domain. On the other side, TM techniques can assist ontologies construction. As Zhou et al. [146] mentioned that applying TM techniques to automatically or semi-automatically construct ontology from unstructured or semi-structured data sources is a promising research direction in AEC domains.

● Application of innovative technologies

TM is a relatively new but fast-growing technology. In last decade, there are various emerging technologies. For instance, deep learning as a rising research direction of machine learning, is a powerful tool to handle space dimensions of large feature in the textual corpus. Some recent researches have integrated deep learning into the text classification [46, 102], and information extraction [116]. Deep learning tends to be a promising direction in TM applications.

Text summarization as one of TM functions, has rarely been applied to the construction industry. Text summarization can transform the full text into a short version with a set of topic sentences. It can provide readers with a summary of enormous documents. Due to the existence of a huge amount of information for users to browse, it is essential to develop automatic summarization technique to provide information efficiently. In future works, text summarization will attract increasing attention from experts in the AEC domain.

● Sentiment analysis

Another potential research direction is sentiment analysis of social media. Since Internet attracts more and more individuals to express their opinions to public affairs in social media, valuable data for administrators, including trends, emotions, and messages related to construction projects could be captured and analyzed for decision making. At the urban level, governors can identify shortcomings of urban construction through sentiment analysis. At the industry level, sentiment analysis can be applied to explore the citizens' attention for the advancement of new technology transformations, such as green building [71] and off-site construction [68]. At the project level, project management team can capture public's opinions at the planning stage or users' opinions at the maintenance and operation stage. Property managers can conduct monthly, weekly, or daily sentiment analysis relying on the constant monitoring, and perform automatic warning and administrative mechanisms. Decision makers should respond to these opinions, especially for negative comments.

Nevertheless, the underlying applications of sentiment analysis in AEC domain have not been completely exploited. In future studies, the utilization of this technology will provide the opportunities for identifying public opinion, attitudes, and emotions economically and rapidly in the construction industry.

7. Conclusion

The construction sector generates an enormous amount of data during the life cycle of a construction project. The majority of the data are from unstructured textual documents. Therefore, TM techniques are introduced to extract potential valuable knowledge from these unstructured textual data in AEC domain. This study comprehensively reviews the applications of TM in AEC industry from the aspects of development history, application areas, challenges, and future trends. 127 journal papers between 2000 and 2021 are selected and a qualitative-quantitative method is applied to analyze these articles. VOSviewer software is used to conduct quantitative analysis, and qualitative analysis is carried out to analyze eight TM application fields in depth. The key findings of this research are explained as follow:

1) During the period from 2000 to 2021, the number of publications is continuously increasing and reached its peak in 2020, among which US and Chinese researchers engaged most actively, compared to researchers from other districts and countries around the world.
2) Nora EI-Gohary from Illinois at Urbana-Champaign is the most productive authors in this research field. In addition, the relationship between authors and influential research teams is also identified.
3) Several important topics and related techniques are identified through keywords analysis. Relationships between these topics and the evolution of main research topics are explored by conducting co-occurrence of keywords. The common methods and tools for TM applications are also identified.
4) Eight TM application fields in the construction industry are identified: safety management, automation compliance checking, building design, method development or improvement, public opinion analysis, framework development, contract management, and others. TM applications in these fields are discussed in-depth.
5) Five challenges of TM applications in AEC domain including the limitations in ACC, limitations in the analysis of public opinion, less utilization of domain knowledge, ignorance of structural characteristics, publicly available datasets. Furthermore, three future directions are proposed: integration of domain knowledge or ontology, application of innovative technologies, and sentiment analysis of social media.

Findings of this study provide the future work with directions and can help scholars understand the current research trend of TM applications in AEC domain. However, a few limitations exist in this study. Specifically, this study only searches five online databases for journal articles from 2000 and 2021, but articles from other databases and conference articles are excluded. In addition, publications in other language are not included in this study. More relevant articles of different databases, publication years and languages can be included in future researches, and different research findings may be captured.

Declarations

Author contribution statement

All authors listed have significantly contributed to the development and the writing of this article.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgement

The authors gratefully acknowledge the funding and support provided by the Ministry of Housing and Urban-Rural Development of the People's Republic of China ("2022-R-022").

References

[1] D.Y. Lee, H.L. Chi, J. Wang, X. Wang, C.S. Park, A linked data system framework for sharing construction defect information using ontologies and BIM environments, Autom. ConStruct. 68 (2016) 102–113.

[2] Đ. Nedeljković, M. Kovačević, Building a construction project key-phrase network from unstructured text documents, J. Comput. Civ. Eng. 31 (6) (2017), 04017058.
[3] L. Soibelman, J. Wu, C. Caldas, I. Brilakis, K.Y. Lin, Management and analysis of unstructured construction data types, Adv. Eng. Inf. 22 (1) (2008) 15–27.
[4] S. Moon, Y. Shin, B.G. Hwang, S. Chi, Document management system using TM for information acquisition of international construction, KSCE J. Civ. Eng. 22 (12) (2018) 4791–4798.
[5] M. Al Qady, A. Kandil, Concept relation extraction from construction documents using natural language processing, J. Construct. Eng. Manag. 136 (3) (2010) 294–302.
[6] R. Feldman, J. Sanger, The TM Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2007.
[7] A. Sun, M. Lachanski, F.J. Fabozzi, Trade the tweet: social media TM and sparse matrix factorization for stock market prediction, Int. Rev. Financ. Anal. 48 (2016) 272–281.
[8] D. Delen, M.D. Crossland, Seeding the survey and analysis of research literature with TM, Expert Syst. Appl. 34 (3) (2008) 1707–1720.
[9] C. Meaney, R. Moineddin, T. Voruganti, M.A. O'Brien, P. Krueger, F. Sullivan, TM describes the use of statistical and epidemiological methods in published medical research, J. Clin. Epidemiol. 74 (2016) 124–132.
[10] R.J. Scherer, S. Reul, Retrieval of project knowledge from heterogeneous AEC documents, in: Computing in Civil and Building Engineering, 2000, pp. 812–819.
[11] M. Chui, Artificial Intelligence the Next Digital Frontier, 47, McKinsey and Company Global Institute, 2017, pp. 3–6.
[12] S. Zhou, S.T. Ng, S.H. Lee, F.J. Xu, Y. Yang, A Domain Knowledge Incorporated TM Approach for Capturing User Needs on BIM Applications, Construction and Architectural Management, Engineering, 2019.
[13] H. Yan, N. Yang, Y. Peng, Y. Ren, Data mining in the construction industry: present status, opportunities, and future trends, Autom. ConStruct. 119 (2020), 103331.
[14] S. Baek, W. Jung, S.H. Han, A critical review of text-based research in construction: data source, analysis method, and implications, Autom. ConStruct. 132 (2021), 103915.
[15] J.S. Kim, B.S. Kim, Analysis of fire-accident factors using big-data analysis method for construction areas, KSCE J. Civ. Eng. 22 (5) (2018) 1535–1543.
[16] F.U. Hassan, T. Le, Automated requirements identification from construction contract documents using natural language processing, J. Leg. Aff. Dispute Resolut. Eng. Constr. 12 (2) (2020), 04520009.
[17] S.M. Weiss, N. Indurkhya, T. Zhang, F. Damerau, TM: Predictive Methods for Analyzing Unstructured Information, Springer Science & Business Media, 2010.
[18] R.A. Erhardt, R. Schneider, C. Blaschke, Status of text-mining techniques applied to biomedical text, Drug Discov. Today 11 (7-8) (2006) 315–325.
[19] VOSviewer, Welcome to VOSviewer, 2020. Retrieved from, https://www.vosviewer.com/.
[20] J. Song, H. Zhang, W. Dong, A review of emerging trends in global PPP research: analysis and visualization, Scientometrics 107 (3) (2016) 1111–1147.
[21] Q. He, G. Wang, L. Luo, Q. Shi, J. Xie, X. Meng, Mapping the managerial areas of Building Information Modeling (BIM) using scientometric analysis, Int. J. Proj. Manag. 35 (4) (2017) 670–685.
[22] J.Y. Park, Z. Nagy, Comprehensive analysis of the relationship between thermal comfort and building control research-A data-driven literature review, Renew. Sustain. Energy Rev. 82 (2018) 2664–2679.
[23] M. Oraee, M.R. Hosseini, E. Papadonikolaki, R. Palliyaguru, M. Arashpour, Collaboration in BIM-based construction networks: a bibliometric-qualitative literature review, Int. J. Proj. Manag. 35 (7) (2017) 1288–1301.
[24] X. Zhao, A scientometric review of global BIM research: analysis and visualization, Autom. ConStruct. 80 (2017) 37–47.
[25] M.R. Hosseini, I. Martek, E.K. Zavadskas, A.A. Aibinu, M. Arashpour, N. Chileshe, Critical evaluation of off-site construction research: a Scientometric analysis, Autom. ConStruct. 87 (2018) 235–247.
[26] L. Shen, H. Yan, H. Fan, Y. Wu, Y. Zhang, An integrated system of TM technique and case-based reasoning (TM-CBR) for supporting green building design, Build. Environ. 124 (2017) 388–401.
[27] H. Fan, H. Li, Retrieving similar cases for alternative dispute resolution in construction accidents using TM techniques, Autom. ConStruct. 34 (2013) 85–91.
[28] H.N. Su, P.C. Lee, Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight, Scientometrics 85 (1) (2010) 65–79.
[29] N.J. Van Eck, L. Waltman, Software survey: VOSviewer, a computer program for bibliometric mapping, Scientometrics 84 (2) (2010) 523–538.
[30] M. Al Qady, A. Kandil, Automatic clustering of construction project documents based on textual similarity, Autom. ConStruct. 42 (2014) 36–49.
[31] T.H. Beach, Y. Rezgui, H. Li, T. Kasim, A rule-based semantic approach for automated regulatory compliance in the construction sector, Expert Syst. Appl. 42 (12) (2015) 5219–5231.
[32] H. Fan, F. Xue, H. Li, Project-based as-needed information retrieval from unstructured AEC documents, J. Manag. Eng. 31 (1) (2015) A4014012.
[33] P. Zhou, N. El-Gohary, Ontology-based automated information extraction from building energy conservation codes, Autom. ConStruct. 74 (2017) 103–117.
[34] C.L. Yeung, C.F. Cheung, W.M. Wang, E. Tsui, A knowledge extraction and representation system for narrative analysis in the construction industry, Expert Syst. Appl. 41 (13) (2014) 5710–5722.
[35] S. Shrestha, S.A. Morshed, N. Pradhananga, X. Lv, Leveraging accident investigation reports as leading indicators of construction safety using text classification, in: Construction Research Congress 2020: Safety, Workforce, and Education, American Society of Civil Engineers, Reston, VA, 2020, November, pp. 490–498.
[36] H. Kim, H.S. Lee, M. Park, B. Chung, S. Hwang, Information retrieval framework for hazard identification in construction, J. Comput. Civ. Eng. 29 (3) (2015), 04014052.
[37] Y. Zou, A. Kiviniemi, S.W. Jones, Retrieving similar cases for construction project risk management using Natural Language Processing techniques, Autom. ConStruct. 80 (2017) 66–76.
[38] T. Kim, S. Chi, Accident case retrieval and analyses: using natural language processing in the construction industry, J. Construct. Eng. Manag. 145 (3) (2019), 04019004.
[39] D. Kifokeris, Y. Xenidis, Application of linguistic clustering to define sources of risks in technical projects, ASCE-ASME J. Risk Uncert. Eng. Sys. Part A: J. Inst. Eng. Civ. Eng. Div. 4 (1) (2018), 04017031.
[40] N.W. Chi, K.Y. Lin, S.H. Hsieh, On effective text classification for supporting job hazard analysis, in: Computing in Civil Engineering, 2013, pp. 613–620.
[41] M.Y. Cheng, D. Kusoemo, R.A. Gosno, TM-based construction site accident classification using hybrid supervised machine learning, Autom. ConStruct. 118 (2020), 103265.
[42] Y.M. Goh, C.U. Ubeynarayana, Construction accident narrative classification: an evaluation of TM techniques, Accid. Anal. Prev. 108 (2017) 122–130.
[43] F. Zhang, H. Fleyeh, X. Wang, M. Lu, Construction site accident analysis using TM and natural language processing techniques, Autom. ConStruct. 99 (2019) 238–248.
[44] H.R. Marucci-Wellman, H.L. Corns, M.R. Lehto, Classifying injury narratives of large administrative databases for surveillance—a practical approach combining machine learning ensembles and human review, Accid. Anal. Prev. 98 (2017) 359–371.
[45] W. Fang, H. Luo, S. Xu, P.E. Love, Z. Lu, C. Ye, Automated text classification of near-misses from safety reports: an improved deep learning approach, Adv. Eng. Inf. 44 (2020), 101060.
[46] S. Mangalathu, H.V. Burton, Deep learning-based classification of earthquake-impacted buildings using textual damage descriptions, Int. J. Disaster Risk Reduc. 36 (2019), 101111.
[47] N.W. Chi, K.Y. Lin, S.H. Hsieh, Using ontology-based text classification to assist Job Hazard Analysis, Adv. Eng. Inf. 28 (4) (2014) 381–394.
[48] M. Martínez-Rojas, R.M. Antolín, F. Salguero-Caparrós, J.C. Rubio-Romero, Management of construction Safety and Health Plans based on automated content analysis, Autom. ConStruct. 120 (2020), 103362.
[49] A.J.P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Automated content analysis for construction safety: a natural language processing system to extract precursors and outcomes from unstructured injury reports, Autom. ConStruct. 62 (2016) 45–56.
[50] Y.G. Choi, K.T. Cho, Analysis of safety management characteristics using network analysis of CEO messages in the construction industry, Sustainability 12 (14) (2020) 5771.
[51] H. Baker, M.R. Hallowell, A.J.P. Tixier, Automatically learning construction injury precursors from text, Autom. ConStruct. 118 (2020), 103145.
[52] D.M. Salama, N.M. El-Gohary, Semantic text classification for supporting automated compliance checking in construction, J. Comput. Civ. Eng. 30 (1) (2016), 04014106.
[53] J. Zhang, N.M. El-Gohary, Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking, J. Comput. Civ. Eng. 30 (2) (2016), 04015014.
[54] C. Eastman, J.M. Lee, Y.S. Jeong, J.K. Lee, Automatic rule-based checking of building designs, Autom. ConStruct. 18 (8) (2009) 1011–1033.
[55] X. Tan, A. Hammad, P. Fazio, Automated code compliance checking for building envelope design, J. Comput. Civ. Eng. 24 (2) (2010) 203–211.
[56] N.O. Nawari, Automating codes conformance, J. Architect. Eng. 18 (4) (2012) 315–323.
[57] J. Melzner, S. Zhang, J. Teizer, H.J. Bargstädt, A case study on automated safety compliance checking to assist fall protection design and planning in building information models, Construct. Manag. Econ. 31 (6) (2013) 661–674.
[58] B.T. Zhong, L.Y. Ding, H.B. Luo, Y. Zhou, Y.Z. Hu, H.M. Hu, Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking, Autom. ConStruct. 28 (2012) 58–70.
[59] S. Li, H. Cai, V.R. Kamat, Integrating natural language processing and spatial reasoning for utility compliance checking, J. Construct. Eng. Manag. 142 (12) (2016), 04016074.
[60] D.A. Salama, N.M. El-Gohary, Automated compliance checking of construction operation plans using a deontology for the construction domain, J. Comput. Civ. Eng. 27 (6) (2013) 681–698.
[61] P. Zhou, N. El-Gohary, Domain-specific hierarchical text classification for supporting automated environmental compliance checking, J. Comput. Civ. Eng. 30 (4) (2016), 04015057.
[62] D.M. Salama, N.M. El-Gohary, Semantic text classification for supporting automated compliance checking in construction, J. Comput. Civ. Eng. 30 (1) (2016), 04014106.
[63] P. Zhou, N. El-Gohary, Ontology-based multilabel text classification of construction regulatory documents, J. Comput. Civ. Eng. 30 (4) (2016), 04015058.
[64] J. Zhang, N.M. El-Gohary, Extending building information models semiautomatically using semantic natural language processing techniques, J. Comput. Civ. Eng. 30 (5) (2016) C4016004.
[65] D. Boyd, Social media: a phenomenon to be analyzed, Social Media+ Society 1 (1) (2015), 2056305115580148.
[66] A. Nikolaidou, P. Papaioannou, Utilizing social media in transport planning and public transit quality: survey of literature, J. Transport. Eng., Part A: Systems 144 (4) (2018), 04018007.

[67] M.A. Kausar, V.S. Dhaka, S.K. Singh, Web crawler: a review, Int. J. Comput. Appl. 63 (2) (2013).
[68] D. Wang, J. Fan, H. Fu, B. Zhang, Research on Optimization of Big Data Construction Engineering Quality Management Based on RNN-LSTM, Complexity, 2018.
[69] R.K. Bakshi, N. Kaur, R. Kaur, G. Kaur, Opinion mining and sentiment analysis, in: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2016, March, pp. 452–455.
[70] P. Ekman, An argument for basic emotions, Cognit. Emot. 6 (3-4) (1992) 169–200.
[71] X. Liu, W. Hu, Attention and sentiment of Chinese public toward green buildings based on Sina Weibo, Sustain. Cities Soc. 44 (2019) 550–558.
[72] T. Hofmann, Probabilistic Latent Semantic Analysis, 2013 arXiv preprint arXiv: 1301.6705.
[73] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[74] G. Ignatow, R. Mihalcea, An Introduction to TM: Research Design, Data Collection, and Analysis, Sage Publications, 2017.
[75] H. Jiang, P. Lin, M. Qiang, Public-opinion sentiment analysis for large hydro projects, J. Construct. Eng. Manag. 142 (2) (2016), 05015013.
[76] R. Bortolini, N. Forcada, Analysis of building maintenance requests using a text mining approach: building services evaluation, Build. Res. Inf. 48 (2) (2020) 207–217.
[77] L. Tang, Y. Zhang, F. Dai, Y. Yoon, Y. Song, R.S. Sharma, Social media data analytics for the US construction industry: preliminary study on Twitter, J. Manag. Eng. 33 (6) (2017), 04017038.
[78] H. Villeneuve, W. O'Brien, Listen to the guests: text-mining Airbnb reviews to explore indoor environmental quality, Build. Environ. 169 (2020), 106555.
[79] Y. Pan, L. Zhang, BIM log mining: exploring design productivity characteristics, Autom. ConStruct. 109 (2020), 102997.
[80] S. Yarmohammadi, R. Pourabolghasem, D. Castro-Lacouture, Mining implicit 3D modeling patterns from unstructured temporal BIM log text data, Autom. ConStruct. 81 (2017) 17–24.
[81] N. Jung, G. Lee, Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning, Adv. Eng. Inf. 41 (2019), 100917.
[82] L. Zhang, B. Ashuri, BIM log mining: discovering social networks, Autom. ConStruct. 91 (2018) 31–43.
[83] X. Zhang, B. Wu, L. Dong, N. Ye, Application of Spark parallelization technology in architectural text classification, J. Comput. Methods Sci. Eng. 18 (4) (2018) 963–976.
[84] M.P. Nepal, S. Staub-French, R. Pottinger, J. Zhang, Ontology-based feature modeling for construction information extraction from a building information model, J. Comput. Civ. Eng. 27 (5) (2013) 555–569.
[85] M. Park, K.W. Lee, H.S. Lee, P. Jiayi, J. Yu, Ontology-based construction knowledge retrieval system, KSCE J. Civ. Eng. 17 (7) (2013) 1654–1663.
[86] P. Demian, R. Fruchter, Measuring relevance in support of design reuse from archives of building product models, J. Comput. Civ. Eng. 19 (2) (2005) 119–136.
[87] X. Liu, B. Akinci, M. Bergés, J.H. Garrett Jr., Domain-specific querying formalisms for retrieving information about HVAC systems, J. Comput. Civ. Eng. 28 (1) (2014) 40–49.
[88] M. Kovacevic, J.Y. Nie, C. Davidson, Providing answers to questions from automatically collected web pages for intelligent decision making in the construction sector, J. Comput. Civ. Eng. 22 (1) (2008) 3–13.
[89] C.H. Caldas, L. Soibelman, Automating hierarchical document classification for construction management information systems, Autom. ConStruct. 12 (4) (2003) 395–406.
[90] C.H. Caldas, L. Soibelman, Implementing automated methods for document classification in construction management information systems, in: Computing in Civil Engineering, 2003, pp. 194–210, 2002.
[91] C.H. Caldas, L. Soibelman, J. Han, Automated classification of construction project documents, J. Comput. Civ. Eng. 16 (4) (2002) 234–243.
[92] M. Gajzler, Text and data mining techniques in aspect of knowledge acquisition for decision support system in construction industry, Technol. Econ. Dev. Econ. (2) (2010) 219–232.
[93] C.L. Yeung, C.F. Cheung, W.M. Wang, E. Tsui, A knowledge extraction and representation system for narrative analysis in the construction industry, Expert Syst. Appl. 41 (13) (2014) 5710–5722.
[94] H.T. Lin, N.W. Chi, S.H. Hsieh, A concept-based information retrieval approach for engineering domain-specific technical documents, Adv. Eng. Inf. 26 (2) (2012) 349–360.
[95] Y. Rezgui, Ontology-centered knowledge management using information retrieval techniques, J. Comput. Civ. Eng. 20 (4) (2006) 261–270.
[96] K. Liu, N. El-Gohary, Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports, Autom. ConStruct. 81 (2017) 313–327.
[97] R. Costa, C. Lima, J. Sarraipa, R. Jardim-Gonçalves, Facilitating knowledge sharing and reuse in building and construction domain: an ontology-based approach, J. Intell. Manuf. 27 (1) (2016) 263–282.
[98] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv. 34 (1) (2002) 1–47.
[99] T. Mahfouz, Unstructured construction document classification model through support vector machine (SVM), in: Computing in Civil Engineering, 2011, pp. 126–133.
[100] M. Al Qady, A. Kandil, Automatic classification of project documents on the basis of text content, J. Comput. Civ. Eng. 29 (3) (2015), 04014043.
[101] Y. Zhu, W. Mao, I. Ahmad, Capturing implicit structures in unstructured content of construction documents, J. Comput. Civ. Eng. 21 (3) (2007) 220–227.
[102] B. Zhong, X. Pan, P.E. Love, L. Ding, W. Fang, Deep learning and network analysis: classifying and visualizing accident narratives in construction, Autom. ConStruct. 113 (2020), 103089.
[103] B.G. Kim, S.I. Park, H.J. Kim, S.H. Lee, Automatic extraction of apparent semantic structure from text contents of a structural calculation document, J. Comput. Civ. Eng. 24 (3) (2010) 313–324.
[104] T. Le, H. David Jeong, NLP-based approach to semantic classification of heterogeneous transportation asset data terminology, J. Comput. Civ. Eng. 31 (6) (2017), 04017057.
[105] P. Demian, P. Balatsoukas, Information retrieval from civil engineering repositories: importance of context and granularity, J. Comput. Civ. Eng. 26 (6) (2012) 727–740.
[106] K.Y. Lin, L. Soibelman, Promoting transactions for A/E/C product information, Autom. ConStruct. 15 (6) (2006) 746–757.
[107] K.Y. Lin, S.H. Hsieh, H.P. Tserng, K.W. Chou, H.T. Lin, C.P. Huang, K.F. Tzeng, Enabling the creation of domain-specific reference collections to support text-based information retrieval experiments in the architecture, engineering and construction industries, Adv. Eng. Inf. 22 (3) (2008) 350–361.
[108] J. Sun, K. Lei, L. Cao, B. Zhong, Y. Wei, J. Li, Z. Yang, Text visualization for construction document information management, Autom. ConStruct. 111 (2020), 103048.
[109] J. McKechnie, S. Shaaban, S. Lockley, Computer assisted processing of large unstructured document sets: a case study in the construction industry, in: Proceedings of the 2001 ACM Symposium on Document Engineering, 2001, November, pp. 11–17.
[110] C.H. Caldas, L. Soibelman, L. Gasser, Methodology for the integration of project documents in model-based information systems, J. Comput. Civ. Eng. 19 (1) (2005) 25–33.
[111] J. Padhy, M. Jagannathan, V.S. Kumar Delhi, Application of Natural Language processing to automatically identify exculpatory clauses in construction contracts, J. Leg. Aff. Dispute Resolut. Eng. Constr. 13 (4) (2021), 04521035.
[112] T. Ko, H.D. Jeong, G. Lee, Natural Language processing–driven model to extract contract change reasons and altered work items for advanced retrieval of change orders, J. Construct. Eng. Manag. 147 (11) (2021), 04021147.
[113] J. Lee, J.S. Yi, J. Son, Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based NLP, J. Comput. Civ. Eng. 33 (3) (2019), 04019003.
[114] T. Le, C. Le, H.D. Jeong, S.B. Gilbert, E. Chukharev-Hudilainen, Requirement text detection from contract packages to support project definition determination, in: Advances in Informatics and Computing in Civil and Construction Engineering, Springer, Cham, 2019, pp. 569–576.
[115] A. Faraji, M. Rashidi, S. Perera, Text mining risk assessment–based model to conduct uncertainty analysis of the general conditions of contract in housing construction projects: case study of the NSW GC21, J. Architect. Eng. 27 (3) (2021), 04021025.
[116] B. Zhong, X. Xing, H. Luo, Q. Zhou, H. Li, T. Rose, W. Fang, Deep learning-based extraction of construction procedural constraints from construction regulations, Adv. Eng. Inf. 43 (2020), 101003.
[117] M. Marzouk, M. Enaba, Text analytics to analyze and monitor construction project contract and correspondence, Autom. ConStruct. 98 (2019) 265–274.
[118] Y. Jallan, E. Brogan, B. Ashuri, C.M. Clevenger, Application of natural language processing and TM to identify patterns in construction-defect litigation cases, J. Leg. Aff. Dispute Resolut. Eng. Constr. 11 (4) (2019), 04519024.
[119] H.B. Gunay, W. Shen, C. Yang, Text-mining building maintenance work orders for component fault frequency, Build. Res. Inf. 47 (5) (2019) 518–533.
[120] S.H. Hong, S.K. Lee, J.H. Yu, Automated management of green building material information using web crawling and ontology, Autom. ConStruct. 102 (2019) 230–244.
[121] Y. Mo, D. Zhao, J. Du, M. Syal, A. Aziz, H. Li, Automated staff assignment for building maintenance using natural language processing, Autom. ConStruct. 113 (2020), 103150.
[122] Y. Mo, D. Zhao, M. Syal, A. Aziz, Construction work plan prediction for facility management using TM, in: Computing in Civil Engineering 2017, 2017, pp. 92–100.
[123] H.S. Ng, A. Toukourou, L. Soibelman, Knowledge discovery in a facility condition assessment database using text clustering, J. Infrastruct. Syst. 12 (1) (2006) 50–59.
[124] B. Zhong, X. Xing, P. Love, X. Wang, H. Luo, Convolutional neural network: deep learning-based classification of building quality problems, Adv. Eng. Inf. 40 (2019) 46–57.
[125] N.W. Chi, K.Y. Lin, N. El-Gohary, S.H. Hsieh, Evaluating the strength of text classification categories for supporting construction field inspection, Autom. ConStruct. 64 (2016) 78–88.
[126] K.Y. Lin, L. Soibelman, Incorporating domain knowledge and information retrieval techniques to develop an architectural/engineering/construction online product search engine, J. Comput. Civ. Eng. 23 (4) (2009) 201–210.
[127] K.Y. Lin, L. Soibelman, Knowledge-assisted retrieval of online product information in architectural/engineering/construction, J. Construct. Eng. Manag. 133 (11) (2007) 871–879.
[128] P. Carrillo, J. Harding, A. Choudhary, Knowledge discovery from post-project reviews, Construct. Manag. Econ. 29 (7) (2011) 713–723.
[129] A.K. Choudhary, J.A. Harding, P. Carrillo, P. Oluikpe, N. Rahman, TM Post Project Reviews to Improve the Construction Project Supply Chain Design, DMIN, 2008, pp. 391–397.

[130] A.K. Choudhary, P.I. Oluikpe, J.A. Harding, P.M. Carrillo, The needs and benefits of TM applications on Post-Project Reviews, Comput. Ind. 60 (9) (2009) 728–740.
[131] N. Ur-Rahman, J.A. Harding, Textual data mining for industrial knowledge management and text classification: a business oriented approach, Expert Syst. Appl. 39 (5) (2012) 4729–4739.
[132] T.P. Williams, J. Gong, Predicting construction cost overruns using TM, numerical data and ensemble classifiers, Autom. ConStruct. 43 (2014) 23–29.
[133] M.R. Hosseini, I. Martek, E. Papadonikolaki, M. Sheikhkhoshkar, S. Banihashemi, M. Arashpour, Viability of the BIM manager enduring as a distinct role: association rule mining of job advertisements, J. Construct. Eng. Manag. 144 (9) (2018), 04018085.
[134] J. Zheng, Q. Wen, M. Qiang, Understanding demand for project manager competences in the construction industry: data mining approach, J. Construct. Eng. Manag. 146 (8) (2020), 04020083.
[135] T. Mahfouz, A. Kandil, S. Davlyatov, Identification of latent legal knowledge in differing site condition (DSC) litigations, Autom. ConStruct. 94 (2018) 104–111.
[136] Y. Lai, C.E. Kontokosta, Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities, Comput. Environ. Urban Syst. 78 (2019), 101383.
[137] A. Goel, L.S. Ganesh, A. Kaur, Social Sustainability Considerations in Construction Project Feasibility Study: a Stakeholder Salience Perspective, Eng. Construct. Architect. Manag. (2020).
[138] B. Zhong, Y. Hei, L. Jiao, H. Luo, J. Tang, Technology frontiers of building-integrated photovoltaics (BIPV): a patent Co-citation analysis, Int. J. Low Carbon Technol. 15 (2) (2020) 241–252.
[139] B.Y. Son, E.B. Lee, Using TM to estimate schedule delay risk of 13 offshore oil and gas EPC case studies during the bidding process, Energies 12 (10) (2019) 1956.
[140] F.C. Pereira, F. Rodrigues, M. Ben-Akiva, Text analysis in incident duration prediction, Transport. Res. C Emerg. Technol. 37 (2013) 177–192.
[141] H. Jiang, P. Lin, M. Qiang, Public-opinion sentiment analysis for large hydro projects, J. Construct. Eng. Manag. 142 (2) (2016), 05015013.
[142] R. Steur, Twitter as a spatio-temporal information source for traffic incident management, Geographi. Info. Manag. App (2014).
[143] M. Chui, D. Farrell, K. Jackson, How government can promote open data and help unleash over $3 trillion in economic value, Innov. Local Gov.: Open Data Info. Tech. 2 (2014).
[144] Y. Guo, Y. Peng, J. Hu, Research on high creative application of case-based reasoning system on engineering design, Comput. Ind. 64 (1) (2013) 90–103.
[145] T.E. El-Diraby, K.F. Kashif, Distributed ontology architecture for knowledge management in highway construction, J. Construct. Eng. Manag. 131 (5) (2005) 591–603.
[146] Z. Zhou, Y.M. Goh, L. Shen, Overview and analysis of ontology studies supporting development of the construction industry, J. Comput. Civ. Eng. 30 (6) (2016), 04016026.

Update

Heliyon

Volume 9, Issue 3, March 2023, Page

DOI: https://doi.org/10.1016/j.heliyon.2023.e14525

Heliyon 9 (2023) e14525

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Corrigendum

Corrigendum to "Overview and analysis of the text mining applications in the construction industry" [Heliyon 8 (12) (December 2022) Article e12088]

Hang Yan [a], Mingxue Ma [b], Ying Wu [c], Hongqin Fan [d], Chao Dong [a,*]

[a] School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan, China
[b] School of Engineering, Design and Built Environment, Western Sydney University, Sydney, Australia
[c] School of Management Science and Real Estate, Chongqing University, Chongqing, China
[d] Department of Building and Real Estate, The Hong Kong Polytechnic University, Hong Kong, China

In the original published version of this article, the affiliation for the author Ying Wu was incorrectly listed as School of Management Science and Real Estate, Chongqing University, Chongqing, China. The correct affiliation for Ying Wu is State Grid Zhongxing Co., Ltd. Lvyuan Branch, Beijing, China. The authors apologize for the errors. Both the HTML and PDF versions of the article have been updated to correct the errors.

DOI of original article: https://doi.org/10.1016/j.heliyon.2022.e12088.
* Corresponding author.
E-mail addresses: chaodong@whut.edu.cn, memorize.cool@163.com (C. Dong).

https://doi.org/10.1016/j.heliyon.2023.e14525
Received 9 March 2023; Accepted 9 March 2023
Available online 15 March 2023