The Fourteenth International Conference on
Computational Structures Technology
23–25 August, 2022 | Montpellier, France

Proceedings of the Fourteenth International Conference on
Computational Structures Technology
Edited by B.H.V. Topping and J. Kruis
Civil-Comp Conferences, Volume 3, Paper 13.4
Civil-Comp Press, Edinburgh, United Kingdom, 2022, doi: 10.4203/ccc.3.13.4

Predominant Research Themes in Using Machine Learning in Structural Health Monitoring

M.Z. Akber and X. Zhang

Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology
Kowloon, Hong Kong

Abstract

Structural health monitoring (SHM) using non-destructive machine learning (ML) based technologies has gained considerable interest in research and industrial communities. Integrating the conventional methods of SHM with novel ML techniques gives robust, sustainable, and promising solutions to SHM. This study presents text mining-based methodology to identify predominant research themes in using ML in SHM. Two analyses are performed on literature data of 375 research studies; (1) co-occurrence analysis of keywords applied on author specified keywords and (2) topic modeling using latent dirichlet allocation (LDA) approach applied on abstracts. The finding shows that the research studies predominantly focus on detecting and classifying structural damages, investigating sensing systems or sensors, and feature extraction and analysis. Moreover, convolutional neural networks and support vector machines are the two mainly used ML algorithms, and bridges, dams, and wind turbines are found as the top three investigated engineering structures. This work can be further extended to include a systematic review of past studies to have an in-depth understanding of using ML in SHM and to find potential contributions and research gaps in the studied area.

Keywords: structural health monitoring, machine learning, predominant research themes, keywords analysis, topic modeling, latent dirichlet allocation.

1 Introduction

Structural health monitoring (SHM) is evolving and encompassing novel, non-destructive and computational techniques such as machine learning. Using ML provides a safe, reliable, and sustainable assessment of the operational health and working conditions of engineering structures [1,2]. The SHM considers key elements: data curation and management, system identification and analysis, condition assessment and monitoring, and decision making or maintenance [3]. Traditionally the accomplishment of these elements involves prescriptive and resource-consuming methods and necessitates repetitive equipment operations. Moreover, implementing these setups may require a temporary shutdown of structures or surrounding works. In contrast, the novel machine learning-based tools provide a data-driven, automated, non-contact, and robust structural health assessment. Therefore, ML-based methods are considered time-saving, cost-effective, and sustainable means for SHM [2-4].

ML models learn data, perform computations, and formulate efficient models that help accomplish practical objectives. In recent years, ML has emerged as a prominent knowledge area to help solve problems in various industries such as education, construction, business intelligence, cybersecurity, manufacturing, healthcare, and material science [5-9]. Owing to the considerable use of ML in SHM, this research aims to identify predominant research themes in this field. Consequently, we followed a text mining-based methodology to analyze literature data.

This study performs co-occurrence analysis of keywords and topic modeling using latent dirichlet allocation to analyze literature data and finds predominant research themes in using ML in SHM. Keywords analysis assists in understanding previous research themes, their relationships, and intellectual orderliness [10,11]. Topic modeling is a natural language processing (NLP) approach that incorporates unsupervised learning of text-related digital data such as scientific articles, educational materials, and online available social media data [12,15]. Topic modeling enables a robust, quantitative, and humanly interpretable sense to identify latent semantic structures in text data. Compared to conventional manual review of research documents, using text-mining-based analysis provides a robust and quantitative way to identify predominant research themes in using ML in SHM.

2 Research methodology

This study follows a text mining-based methodology to identify predominant research themes in using ML in SHM, as shown in the Figure 1. Literature data is sourced from a well-known web of science research database [16]. Based on the objective of this

study, the two phrases “machine learning” and “Structural Health Monitoring” are used to search the relevant studies. The search query was designed to include only two document types: journal articles and proceedings papers. Moreover, another filter was applied to have only those documents available in English. Consequently, 375 studies are found based on the search query, and literature data related to these studies are downloaded for further analysis and modeling. After collecting the scientometric data related to 375 studies, two types of analysis are performed: (1) co-occurrence analysis of keywords and (2) topic modeling using the latent dirichlet allocation (LDA) method.

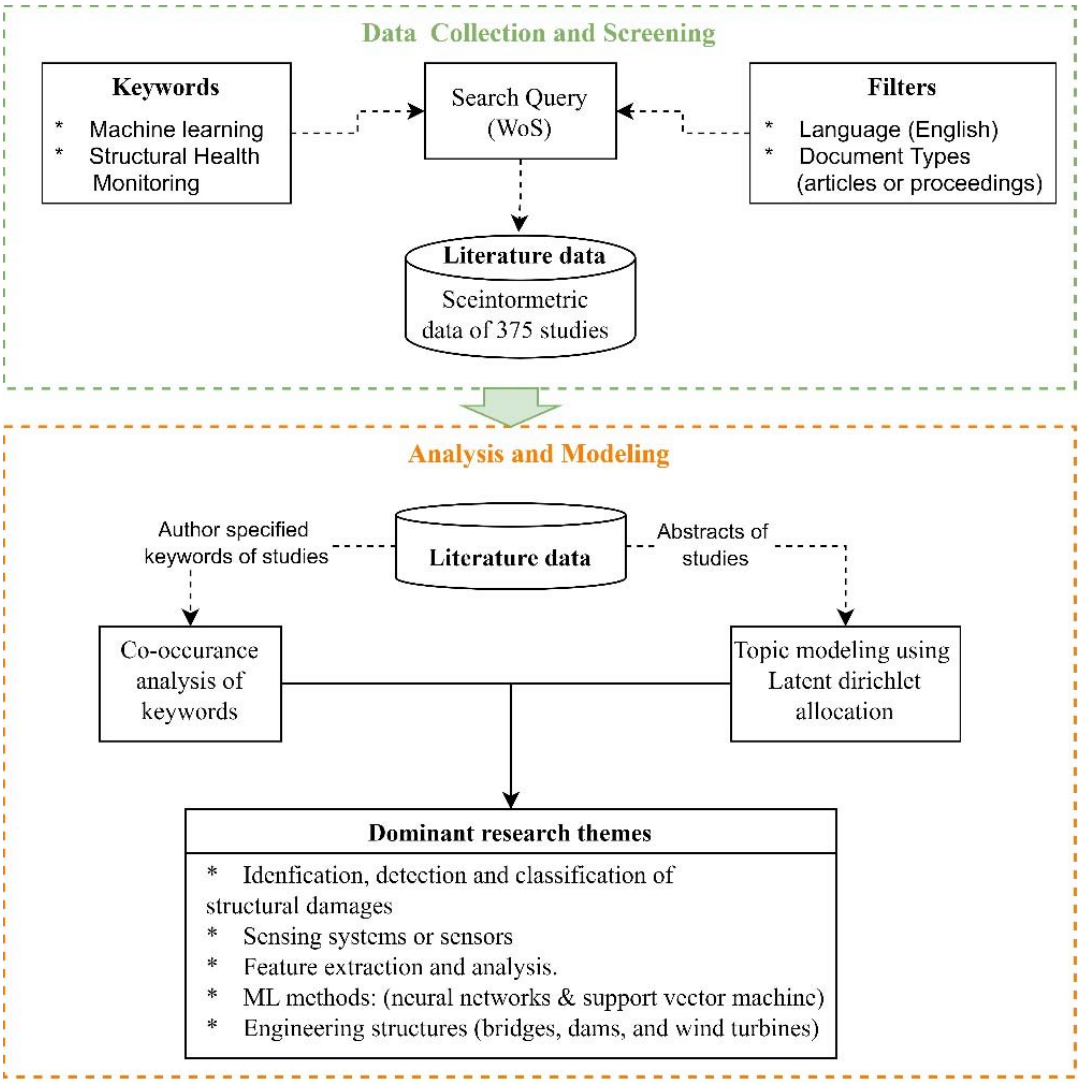

Figure 1 Research methodology.

The keywords analysis is performed using a freely available VOSviewer software widely applied to perform science mapping of literature data. Co-occurrence analysis

is performed on author-specified keywords with a threshold of "minimum number of occurrences of the keyword" = 4. Moreover, to have consistency in keyword mapping, some modifications are made. For example, the two searching keywords 'machine learning' and 'structural health monitoring' are not included. Consider another example where out of two keywords: 'support vector machine' and 'support vector machines', only the keyword of 'support vector machine' is kept in the network since it has the most occurrence.

LDA model is used to perform topic modeling of abstracts of studies. LDA can rapidly extract information from text documents and have a basic idea that each document represents a finite mixture of topics, and each topic represents the finite distribution of words. Analyzing these distributions, the most suitable words for each topic are identified, and a humanly interpretable sense of topics is achieved [13]. LDA is performed on given data for a fixed number of topics, and the probability of an i^{th} word specific to a document is given by the following equation [17].

$$p(w_i) = \sum_{j=1}^{N} p(w_i|z_i = j)p(z_i = j) \tag{1}$$

where $p(w_i|z_i = j)$ is the probability of the ith word conditional to the jth topic and $p(z_i = j)$ is the probability of selecting a word from topic j for a given document.

The LDA is performed on abstracts of 375 shortlisted studies. Before LDA, data preprocessing and cleaning were performed. The abstracts were cleaned from the redundant text that includes phrasings related to Copywrite statements, trademarks publishing, and information. Moreover, punctuations, symbols, and numbers were also removed from text data. After data cleaning, the data were preprocessed to bring consistency in terms. For example, the bigram and trigram were formed for the terms based on two and three words. Then, the LDA is tuned using this final processed data, and the optimum value of the number of topics is determined. Finally, the top terms are extracted related to topics.

3 Results and discussion

The results of co-occurrence analysis of keywords are described using a science mapping network. The science mapping network consists of two main elements: (1)

nodes that represent the keywords and (2) edges that represent the strength of association between keywords [18].

Figure 2 shows the finalized network that contains 37 nodes and 104 edges. The keywords with a larger node size indicate a particular focus of the ML in SHM. For example, the top 10 keywords are ‘damage detection’, ‘deep learning’, ‘support vector machine’, ‘pattern recognition’, ‘damage identification’, ‘neural networks’, ‘convolutional neural networks’, ‘computer vision’, ‘acoustic emission’, and ‘damage classification’. Moreover, strong connections between two research areas are identified from thicker edges between pairs of keywords, for example, ‘deep learning’ – ‘computer vision’, ‘deep learning’ – ‘damage identification’, ‘damage detection’ – ‘pattern recognition’ and ‘domain adaptation’ – ‘transformation learning’.

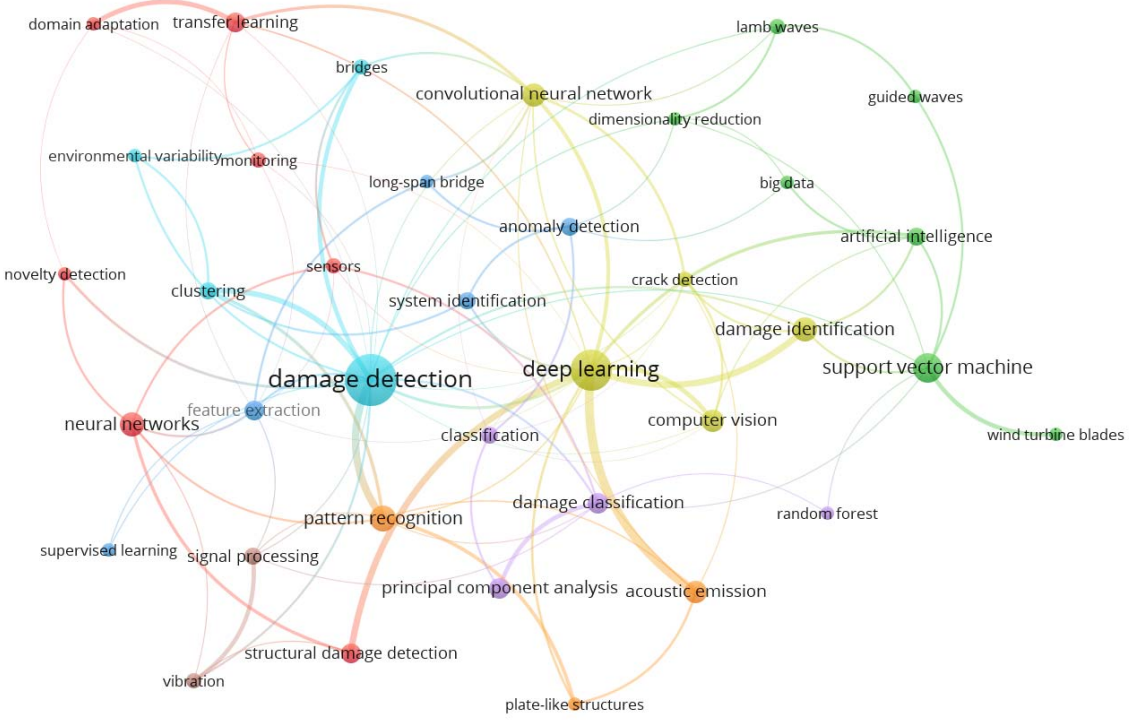

Figure 2 Science mapping network of co-occurrence author specified keywords.

A second analysis performed to identify research themes in using ML in SHM is topic modeling using LDA. The LDA model is optimized for coherence score to choose the best number of topics from 1 to 20 (see Figure 3). Coherence gives semantic similarity between high-scoring terms in the topic; the higher the coherence

higher will be a humanly interpretable sense within the terms of a topic [13,19]. The maximum coherence is seen for k = 16.

The top ten terms related to each topic are shown in Figure 4. For example, the top terms in topic 1 are ‘sensor’, ‘strain’, ‘antenna’ , ‘system’ and ‘building’. Topic 2 has top terms of ‘deep learning’, ‘features’, ‘framework’ and ‘convolutional neural networks’. Using LDA results, a distinct theme can be assigned to some topics; for example, topics 1 and 4 may be related to strain sensor use, topics 3 and 14 to SHM in bridge structures, and topics 9 and 13 to feature extraction and classification. However, due to unsupervised learning and the inherent definition of LDA, the research themes are not evident for some topics. Moreover, some topics might be related to more than one theme, and others might be sub-themes of a large theme. Nonetheless, the ultimate labeling and classification of topics are subjective to the researcher's understanding and subject matter expertise.

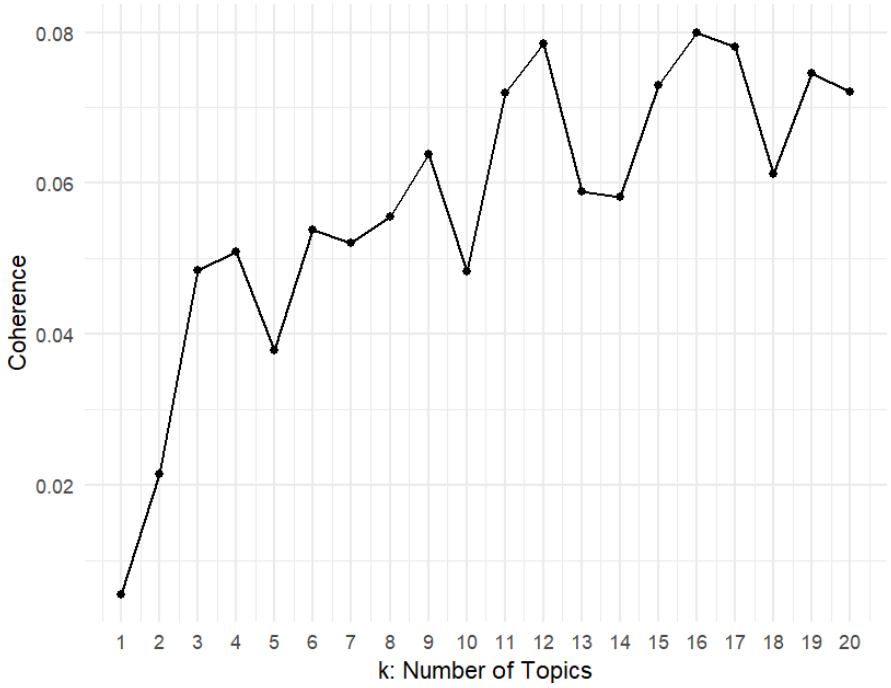

Figure 3 LDA tuning results for the number of topics.

Based on the results of two text-mining based analyses, the following key findings can be drawn: (1) A particular focus on the use of ML in SHM is found in the identification, detection, classification, and monitoring of structural damages, (2) use of sensors and sensing systems are potentially investigated and used together with ML base modeling in SHM, (3) another significance use of ML is seen in features extraction, analysis, classification, and selection, (4) the top ML methods used in SHM includes support vector machine, neural networks, and deep learning based

convolutional neural networks, (5) top engineering structures that are investigated include bridges, dam, and wind turbine.

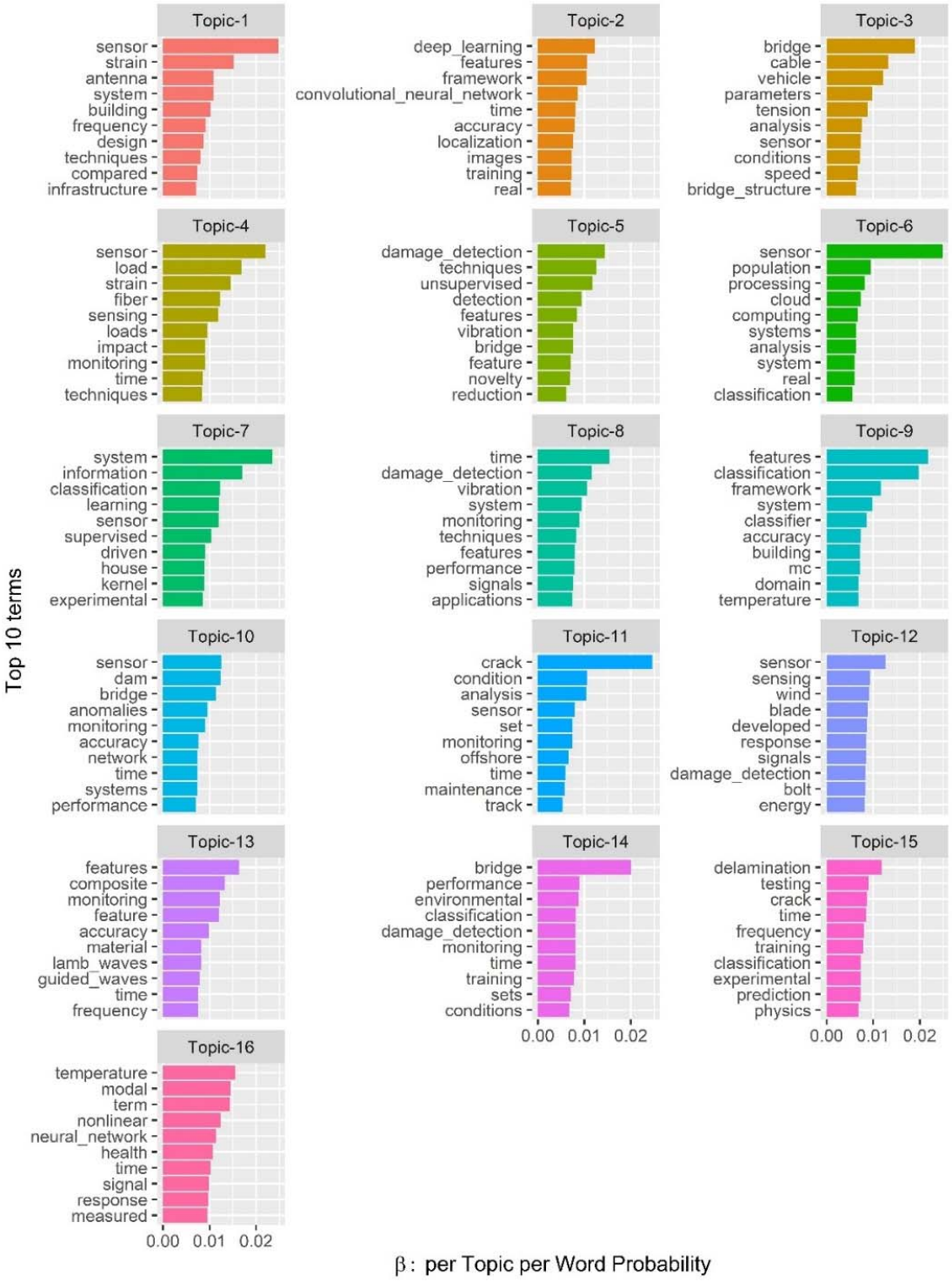

Figure 4 LDA topic modeling results: Top 10 terms in 16 topics.

4 Conclusions and future works

This study uses a text-mining based methodology to find predominant research themes on the use of machine learning (ML) in structural health monitoring (SHM). Scientometric data related to 375 research studies are collected from the web of science (WoS) database and subjected to keywords analysis and topic modeling. Co-occurrence analysis of keywords identifies 37 distinct research keywords. Moreover, latent dirichlet allocation performed on abstracts of studies identifies 16 research topics. The emphasized research themes in using ML in SHM include detection and classification of structural damages, investigation of sensing systems or sensors, and feature extraction and analysis. Moreover, neural networks and support vector machines are the two most used ML algorithms. Among the neural networks, deep learning-based convolutional neural networks are potentially applied to provide a computer vision-based damage detection of engineering structures. Moreover, three engineering structures predominantly investigated using ML include bridges, dams, and wind turbines.

This research identifies predominant research themes in using ML in SHM; however, the comprehensive discussions about the scope and methods related to each theme are not included. Future studies can perform detailed content analysis and systematic literature review and discuss the potential contributions and findings of previous studies. Moreover, this study only used two keywords to design the search query when finding the relevant literature document; future studies may also incorporate overlooked or other specific keywords to expand or broaden the review of past studies. Nonetheless, the text-mining based approach used in this study can be followed to identify predominant research themes in other knowledge areas.

References

[1] F. Kang, X. Liu, and J. Li, “Temperature effect modeling in structural health monitoring of concrete dams using kernel extreme learning machines”, Structural Health Monitoring, 19, 4, 987–1002, 2020.

[2] B. J. Perry, Y. Guo, and H. N. Mahmoud, “Automated site-specific assessment of steel structures through integrating machine learning and fracture mechanics”, Automation in Construction, 133, 2022.

[3] S. Sony, S. Laventure, and A. S. Health, “A literature review of next-generation smart sensing technology in structural health monitoring”, Structural Control and Health Monitoring, 26, 3, 2321, 2019.

[4] C. Dong and F. Catbas, “A review of computer vision–based structural health monitoring at local and global levels”, Structural Health Monitoring, 20, 2, 692–743, 2021.

[5] M. Eminagaoglu and S. Eren, “Implementation and comparison of machine learning classifiers for information security risk analysis of a human resources department”, “2010 International Conference on Computer Information Systems and Industrial Management Applications, CISIM”,2010.

[6] B. Jeong, H. Cho, J. Kim, S.K. Kwon, S.W. Hong, C.S. Lee, T.Y. Kim, M.S. Park, S. Hong, T.Y Heo,“Comparison between statistical models and machine learning methods on classification for highly imbalanced multiclass kidney data”, Diagnostics, 10, 6, 415, 2020.

[7] X. Zhang, M. Z. Akber, and W. Zheng, “Prediction of seven-day compressive strength of field concrete”, Construction and Building Materials, 305, 124604, 2021.

[8] Y. Liu, T. Zhao, W. Ju, and S. Shi, “Materials discovery and design using machine learning”, Journal of Materiomics, 3, 3, 159–177, 2017.

[9] D. C. Feng, Z. T. Liu, X. D. Wang, Y. Chen, J.Q. Chang, D.F. Wei, Z.M. Jiang, “Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach”, Construction and Building Materials, 230, 117000, 2020.

[10] R. Akram, M. J. Thaheem, A. R. Nasir, T. H. Ali, and S. Khan, “Exploring the role of building information modeling in construction safety through science mapping”, Safety Science, 120, 456–470, 2019.

[11] A. Darko, A. P. C. Chan, X. Huo, and D. G. Owusu-Manu, “A scientometric analysis and visualization of global green building research”, Building and Environment, 149, 501–511, 2019.

[12] T. M. Atkinson, “Latent dirichlet allocation in discovering goals in patients undergoing bladder cancer surgery”, “Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018” 2019.

[13] L. Muchene and W. Safari, “Two-stage topic modelling of scientific publications: A case study of University of Nairobi, Kenya”, PLoS ONE, 16, 1 January, e0243208, 2021.

[14] D. Blei, L. Carin, and D. Dunson, “Probabilistic topic models”, IEEE Signal Processing Magazine, 2010.

[15] J. Silge and D. Robinson, *Text mining with R: A tidy approach*, “ O’Reilly Media, Inc.,” 2017.

[16] A. Aghaei Chadegani, H. Salehi, M. M. Md Yunus, H. Farhadi, M. Fooladi, M. Farhadi, N. Ale Ebrahim, “A comparison between two main academic literature collections: Web of science and scopus databases”, Asian Social Science, 9, 5, 18–26, 2013.

[17] Y. Li, B. Rapkin, T. M. Atkinson, E. Schofield, and B. H. Bochner, “Leveraging Latent Dirichlet Allocation in processing free-text personal goals

among patients undergoing bladder cancer surgery", Quality of Life Research, 28, 6, 1441–1455, 2019.

[18] N. J. Van Eck and L. Waltman, "Manual for VOSviwer version 1.6.10", CWTS Meaningful metrics, 2019.

[19] C. Fay, " Text Mining with R : A Tidy Approach", Journal of Statistical Software, 2018.