Preprints.org

Article

Not peer-reviewed version

Using Data Mining and Machine Learning for Knowledge Discovery in Publications on Information Management and Data Insights

José Eduardo Ricciardi Favaretto * and Eduardo de Rezende Francisco

Posted Date: 16 November 2026

Keywords: knowledge discovery; social network analysis; text mining; clustering; topic modeling; data analysis

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Using Data Mining and Machine Learning for Knowledge Discovery in Publications on Information Management and Data Insights

José Eduardo Ricciardi Favaretto [1,2],* and Eduardo de Rezende Francisco [2]

[1] Graduate Program in Controllership, Finance, and Management Technologies (PPGCFTG), Center for Social and Applied Sciences (CCSA) - Mackenzie Business School, Mackenzie Presbyterian University (UPM), Sao Paulo, SP, Brazil

[2] Getulio Vargas Foundation (FGV EAESP), Center for Studies in Analytics and Security Policy (FGV Analytics), Sao Paulo, SP, Brazil

* Correspondence: jose.favaretto@mackenzie.br

Abstract

In an era of growing data volumes, advances in data storage and processing technologies have created a landscape in which organizations — whether commercial, government, or educational — increasingly request analytical methods that convert raw data into valuable, actionable insights and knowledge. This research employed Data Mining (DM) techniques — including text mining and social network analysis — and Machine Learning (ML) methods, specifically hierarchical agglomerative and K-means clustering, and Latent Dirichlet Allocation (LDA) for topic modeling, to support knowledge discovery across 287 scientific publications focused on information management and data insights. Main findings include: artificial intelligence and machine learning were the most common bigrams in titles; five clusters were identified, with cluster C1 containing the five most cited articles; LDA modeling identified six topics, with Topic 6, Machine Learning/Deep Learning and Predictive Modeling, containing 59 articles on technical and algorithmic aspects. This work could motivate researchers in Data, Information, and Management to deepen their understanding of the field, guide strategic academic development, and promote new knowledge and scientific advances within the discipline.

Keywords: knowledge discovery; social network analysis; text mining; clustering; topic modeling; data analysis

1. Introduction

The field of Data, Information, and Management, as well as related fields such as Data Sciences, Information Sciences, Management Sciences, Decision Sciences, Decision Analysis, Decision Intelligence, and Decision Support Systems, among others, balances practical application with the use of theoretical knowledge, which is reflected in the field's development and reinforces its identity and interdisciplinary and multidisciplinary nature.

In an era of growing data volumes, advances in data storage and processing technologies have created a landscape in which organizations — whether commercial, government, or educational — increasingly request analytical methods that convert raw data into valuable, actionable insights and knowledge.

An editorial advocates a paradigm shift toward hybrid analytical architectures (Limongi et al., 2025), arguing that integrating complementary analytical methods yields insights that substantially exceed those produced by individual techniques alone ("1+1 > 2").

In this context, highlight two key research interests in these fields, with a focus on data-driven knowledge discovery to support decision-making, including Data Mining (DM) (López-Meneses et

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

al., 2025; Shaik et al., 2023) and Machine Learning (ML) (Biju et al., 2024; Husna et al., 2024), which are often seen as synonyms but have distinct objectives, approaches, and foundations. While DM seeks to identify and extract hidden patterns, rules, and summaries in data, discover relationships within datasets, and focus on exploration and descriptive analysis, and it can also serve as a preliminary step for insights or data preparation for models, ML is highly automated, with algorithms that learn from data and adjust their parameters to improve performance on new data, focusing on generalization.

DM enables researchers to identify various patterns, with the selection of a method depending on the data type and analysis objectives (Ge et al., 2024). Artificial Intelligence (AI) and ML algorithms (Barbierato et al., 2025; Biju et al., 2024) are seen as an essential component for identifying complex patterns and predicting trends by enabling data-driven decision-making and predictive modeling (Yin et al., 2025).

Other studies have confirmed the importance of using DM and ML in various contexts, including educational data mining (EDM), where several types of algorithms at different levels of machine learning (Ordoñez-Avila et al., 2023) could contribute to the development of predictive models; Gibert et al. (2018) have proposed a new methodology for linking the characteristics of a given environmental problem to a variety of DM techniques, providing usage guidelines for non-expert users to help them decide which technique is most appropriate for solving their specific problem; and Dogan & Birant (2021) presented a comprehensive literature review of the use of ML and DM techniques and tools – key approaches and algorithms – that can be useful in addressing the challenges in manufacturing organizations, and outlined the main steps in the Knowledge Discovery in Databases (KDD) to be followed by these organizations. Additionally, a study by Rejeb et al. (2025), which examined 10,176 articles using analytical approaches to outline thematic evolution and investigate knowledge transmission pathways in Industry 4.0, led to the development of a strategic map identifying four distinct sub-periods, each characterized by a specific thematic focus.

These studies reinforce the rationale for using data from scientific publications as an alternative for discovering new knowledge and mapping the progress of a specific field.

The general objective of this research was to apply a variety of DM analytical techniques — including text mining and social network analysis — and ML methods, specifically hierarchical agglomerative clustering, K-means clustering, and Latent Dirichlet Allocation (LDA) for topic modeling, to support knowledge discovery in 287 scientific publications focused on information management and data insights. The research question guiding this work was: How can DM and ML analytical techniques support knowledge discovery from this collection of scientific articles?

The contributions and results derived from these analytical approaches facilitate a deeper understanding of research dynamics and their influences, identify prevailing trends and emerging issues, and offer a comprehensive overview of prominent areas, thereby aiding the entire scientific community.

2. Related Studies and Conceptualizations

2.1. Data Mining, Machine Learning, and Knowledge Discovery

In the initial stages, there was limited consensus regarding what the term DM encompassed, and it can be argued that, in some sense, this remains the case (Coenen, 2011). Although the scientific community had already been using the term DM in the 1980s, researcher Gregory Piatetsky-Shapiro coined the term Knowledge Discovery in Databases (KDD) for the first workshop on this theme (KDD-1989), held in Detroit, USA. However, it was from the 1990s onward that the term DM use intensified – while the term KDD became popular in the AI and ML community, the term DM became much more popular in the press and among businesspeople (Piatetsky-Shapiro, 2000).

According Fayyad et al. (1996) KDD refers to the overall process, including all data preparation and post-processing, while DM refers to the step or sub-process [in a KDD cycle] of applying algorithms to the clean data (i.e., data already prepared for exploration and analysis). These authors

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

also reinforce that DM is "a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data". Even today, the terms DM and KDD are frequently used interchangeably or as synonyms; however, technically, one constitutes a sub-process of the other.

On the other hand, although DM is also frequently used interchangeably with ML in less technical settings, a closer examination shows clear ontological and functional differences. DM is a process that uses techniques from statistics, mathematics, and AI to discover hidden or otherwise non-obvious patterns, relationships, properties, and insights in data. However, since DM, by itself, is generally a manual or semi-automatic process with static adaptation capability (limited to the analyzed data) and is not necessarily aimed at clarifying a research or business problem, it may seek to discover what is in the data without making assumptions about the results to be obtained. DM focuses primarily on knowledge discovery and aims to yield outcomes such as rules, patterns, and summaries. From a data perspective, DM complements a set of processes (cleaning and selection, integration/pre-processing, and transformation) that are considered to have already been executed.

ML can be considered a subfield of the AI discipline (Barbierato et al., 2025; Biju et al., 2024; Favaretto, 2024b), due to the highly automated nature of its processes. This subfield focuses on developing algorithms and models that learn from data. In other words, with autonomous growth capacity, the models adjust their parameters as new data are introduced, demonstrating dynamic adaptation and the ability to improve their performance through learning and experience. ML focuses primarily on different types of learning (unsupervised, supervised, and reinforcement learning) and on generalization (i.e., generating models that perform well on new, previously unseen data). From a data perspective, ML algorithms aim to improve their performance on specific tasks without being explicitly programmed to do so.

However, there is a perception that this boundary of interdisciplinarity among the three terms is fluid: ML provides many of the techniques that can be used in the DM step of KDD, while DM can be employed as a preliminary stage to identify insights or to prepare the data adequately for use with ML, which reflects benefits in the execution and accuracy of ML models.

Despite advances in cloud computing and ML making analytics widely accessible, companies still face key non-technical bottlenecks: data preparation consumes roughly 80% of management effort, and experts in both business strategy and data science are scarce (Francisco et al., 2019). Data analysis and information management are not merely operational supports for organizations; they are indispensable strategic assets that ground decision-making in evidence and sustain operational competitive advantage (Favaretto, 2015; Favaretto & Meirelles, 2015).

The combined application of DM and ML, guided by the KDD framework for knowledge discovery, which brings together a set of processes beyond DM, should be an iterative process with feedback cycles, aimed at identifying insights from existing and available data and transforming them into new, practical, valuable, and useful knowledge for decision-making.

The multidisciplinary nature of DM and ML demonstrates their potential for use across various areas of knowledge, such as medical (Ge et al., 2024), supply chain (Gelastopoulos & Keramydas, 2025; Yin et al., 2025), education (López-Meneses et al., 2025), manufacturing (Dogan & Birant, 2021) among others, or even through its variations in usage, such as opinion mining (Pathan & Prakash, 2022; Shaik et al., 2023), process mining (Abrehdari, 2025), literature mining (Hooton et al., 2020; Martins et al., 2023), or even visual insight mining (Lian et al., 2025).

2.2. Clustering

Clustering, or Cluster Analysis, is a well-known analytical technique widely used in Business and Management studies (V. Pereira & Bamel, 2021). *Agglomerative hierarchical clustering* (S. Altuntas et al., 2022; Lu et al., 2008) is a method that iteratively merges pairs of clusters to enhance intra-cluster similarity while reducing inter-cluster similarity. Initially, each observation forms its own cluster; these are then merged in pairs until all observations form a single cluster.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

This clustering method produces a tree-like diagram called a *Dendrogram* (Dinh et al., 2025) which shows how layers are merged during cluster formation. Analyzing the distances in the diagram helps determine whether the clusters align with the research problem and guides the selection of the optimal number of clusters.

The partitional k-means clustering method is popular due to its simplicity and effective results, but it can be significantly affected by outlier observations (Jain, 2010; Sinaga & Yang, 2020; Steinley, 2006). Once the value of "k" is set, the observations are randomly assigned to "k" groups and grouped together into cohesive clusters.

The use of the cluster analysis approach is identified in several studies: to categorize 80 of Brazil's largest agribusiness companies based on financial performance and liquidity indicators, identifying patterns and dividing the companies into four distinct groups ("Profitable," "Shrinking," "Healthy," and "Diversified") (V. R. Pereira et al., 2026); in a data-driven approach study, with the k-means clustering technique and datasets elicited under a Picture Fuzzy Set (PFS) environment (Ybañez et al., 2023); in exploring and identifying dominant factors (groups of indicators) separated into clusters to ensure sustainability (Suha & Sanam, 2023); in discovering knowledge structures and clusters in the digital marketing field of study (Saheb et al., 2021); and in identifying groups of articles that were selected on the topic of Big Data Analytics in sustainable tourism (Agrawal et al., 2022).

Hierarchical agglomerative clustering and partitional k-means clustering may not produce consistent outcomes, even when analyzing identical datasets.

2.3. Text Mining

Text Mining techniques (Gelastopoulos & Keramydas, 2025; Spada et al., 2025) enable the identification of terms and words across texts, patterns, and possible trends in study themes that would be difficult to detect when examining each document manually. In a scientific context, these techniques can extract and identify the most frequent terms (n-grams) (e.g., unigrams, bigrams, trigrams) mentioned by an academic community in the text of their scientific publications. A unigram is a single-word term, a bigram is a two-word term, and a trigram is a three-word term. By identifying n-grams in a set of scientific articles, researchers can map the main themes, most frequent topics, and concepts discussed in these publications, deepen understanding of the content, and create relationships between ideas to inform further research.

Rather than relying on a researcher's subjective interpretation when reading and manually categorizing 'snippets' of article content—an approach that can introduce human bias—Text Mining methods use transparent, analytical criteria supported by reproducible algorithms.

A study adopted natural language processing summarization techniques to find, in contemporary times, the newly emerging areas of management supported by big data (Kushwaha et al., 2021). Another study, which applied text mining to scientific articles in the field of services and management, identified areas that were either explored or left unexplored (Kumar et al., 2021).

Analysis of results from the application of *Text Mining* techniques (Usai et al., 2018) can indicate topics (themes) and emerging research areas, which are essential for understanding the development, relevance, and structure of scientific fields, and for directing future studies and collaborations among researchers and institutions.

2.4. Social Network Analysis

In the scientific context, *Social Network Analysis* (SNA) (Singh et al., 2024) examines relationships and interactions among scientific publications (articles), words, researchers, co-authors, and educational institutions. Using concepts, methods, and techniques from network theory, this type of analysis can detect emerging topic networks (*co-occurrence* of words) (Gaviria-Marin et al., 2019) or identify more influential collaboration networks (co-authorship of articles), allowing the knowledge generated within a field of study, a research theme, or even extracted from a collection of scientific journal publications to be mapped and explored by a scientific community.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

When interpreting the visual results of a social network's formation, the so-called *Centrality Measures* (Degree Centrality, Betweenness Centrality, Closeness Centrality, Eigen Centrality) must be used (Abbasi et al., 2012), each offering a different perspective when evaluating the *nodes* of a network - the points that connect an element or entity within the network that is studied, which could be a researcher, an institution, or even words. These *nodes* are interconnected by *edges* (or *links*); such interconnections must be characterized by a type of relationship, for example, the co-occurrence of words/terms in publications or collaborations between authors of publications, among other possibilities.

The literature reports papers that have explored the use of SNA to model user interactions on Twitter as a weighted, targeted network (Logan et al., 2023); in construction project management (CPM) research (Zheng et al., 2016); in six fields of business and management studies (Project management, Risk management, Tourism management, Supply Chain management, Knowledge management, Technology and Innovation management), when analyzing the co-occurrences of keywords in scientific publications between 2001 and 2020 (Anugerah et al., 2022); and on tourism sector data from regions in Türkiye to make the right quarantine decisions in the COVID-19 pandemic (F. Altuntas et al., 2022).

Word *co-occurrence* networks (Gaviria-Marin et al., 2019), or *co-word* networks, refer to the same concept and describe a way to form a network of connections between words that frequently occur together or are identified as available together or nearby in the same block of text, which could be the titles of papers, the lists of keywords provided by the authors, or even a collection of abstracts. This SNA technique uncovers hidden relationships and research themes and maps intellectual structures across disciplines in text datasets.

2.5. Topic Modeling

Topic Modeling (Hankar et al., 2025; Ma et al., 2021; Vayansky & Kumar, 2020) is an analytical technique in machine learning and natural language processing that extracts groups of words from textual data and is commonly used to identify the hidden thematic structure (latent topics) in a collection of documents.

Because it is an unsupervised learning technique, it does not require user-generated labels in the training data (labeled data), unlike supervised classification tasks, and it allows the automatic discovery of thematic structures useful for exploratory analysis of large volumes of unstructured text.

The best-known *Topic Modeling* algorithms include Latent Semantic Analysis (LSA), presented in papers from around the 1990s (Deerwester et al., 1990; Dumais et al., 1988), which was one of the first automated techniques for topic extraction and modeling, and Latent Dirichlet Allocation (LDA) (Blei, 2012; Blei et al., 2003), one of the probabilistic implementations of unsupervised learning that determines the underlying topics of a collection of text documents (discovers latent patterns in textual data).

LDA operates on each document in a text collection that has been previously converted to numerical representations for mathematical processing, using a procedure called Count Vectorization or Bag-of-Words (BoW). This means the algorithm ignores word order and context, focusing solely on word frequency and co-occurrence within each document.

Literature reports several studies that employed the analytical technique of topic modeling across various contexts: the authors Lee & Kang (2018) utilized 11,693 scientific articles from 1997 to 2016 in the field of Technology and Innovation Management (TIM) and applied the LDA model to identify fifty topics within TIM research, providing concise overviews of the top ten most prominent topics; Sonk & Tunger (2024) addressed the question of how a combination of methods (text mining and topic modeling) could contribute to monitoring trends in a set of 4,415 abstracts from scientific publications listed in the Web of Science (WoS) from 1991 to 2021; and a study of Bongini et al. (2022) analyzed the content of white papers using a sample of 188 Security token offerings (STOs) from 2017 to 2021, applying LDA topic modeling to identify topics and themes and determine which topics were related to campaign success.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

The exploration of *Topic Modeling* analytical techniques should be expanded by using different existing methods, considering the various types of relationships and constraints present in the datasets used in research, and prioritizing those that most effectively meet the researcher's specific requirements (Vayansky & Kumar, 2020).

3. Method, results, and analysis

3.1. Research Dataset

This research used data from 287 scientific articles published in the International Journal of Information Management Data Insights (IJIMDI), ISSN 2667-0968, a journal published by Elsevier since 2021 (ScienceDirect, 2026) which is listed in the Q1 category of the Scopus indexer (Scopus, 2024) and is included in six subject areas, two of which are "Business, Management and Accounting: Management Information Systems" and "Computer Science: Information Systems".

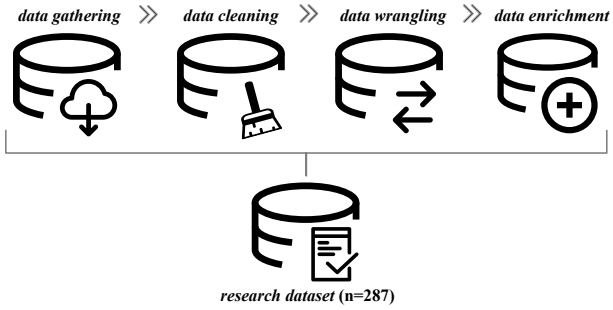


Figure 1. The four phases (stages) carried out in the process of collecting and organizing the research dataset.

Through consultation and data extraction from various sources (*data gathering*), meticulous work was required to organize, clean, and process the data (*data cleaning*), and this was further complemented by transformation, integration, and data enrichment (*data wrangling* / *data enrichment*), resulting in a dataset suitable for exploration during the research – Figure 1.

During the *data gathering* stage of this research, raw data were obtained from three sources: 1) from the Crossref repository (Crossref, 2024), which provided most of the metadata for each article published in the journal, such as title, year of publication, and the quantity of references for each publication (*qty_ref*), among others; 2) from the journal's own website (ScienceDirect, 2026), where complementary data were obtained, such as *keywords* provided by the authors, and the texts of the *abstracts* for each publication; 3) from the Scopus journal indexer (Scopus, 2024) where the number of citations each article received on this platform was obtained. With this data, the researcher conducted the data-cleaning stage to ensure data quality (accuracy and consistency) by correcting errors, removing duplicate records, identifying and handling missing values, and addressing inconsistencies and formatting issues.

Next, the *data wrangling* stage was executed, which consisted of data transformation, for example, generating a numeric variable to store the total number of characters existing in article titles including white spaces (*nc_tit*), generating another numeric variable containing the count of the number of co-authors of who collaborated with each published article (*coaut_n*), etc., in addition to the integration (merging) of data obtained from those three different sources (Crossref repository, journal website, and Scopus indexer), ending with the *data enrichment* stage, which focused on improving the dataset by incorporating the total citations of articles obtained from the Scopus indexer (variable *cita_tot* and other variables containing the subtotals of the number of citations for each year), thus resulting in the structuring and compilation of an integrated dataset that guided the research.

Based on the exploration of this compiled dataset, it was possible to perform several additional data extractions and conversions, mostly using *R software* (R Core Team, 2026) and *RStudio* software

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

(Posit Team, 2026) thus allowing the generation of additional information and the identification of several findings, as detailed and reported in the five subsequent subsections: 3.2) Descriptive Statistics; 3.3) Cluster Analysis; 3.4) Text Mining; 3.5) Social Network Analysis; and 3.6) Topic Modeling.

3.2. Descriptive Statistics

It was also possible, through the exploration of the research dataset and the Scopus indexer (Scopus, 2024), to conduct analyses using descriptive statistics. Figures 2, 3, and 4 and Table 1 present the findings from these analyses.

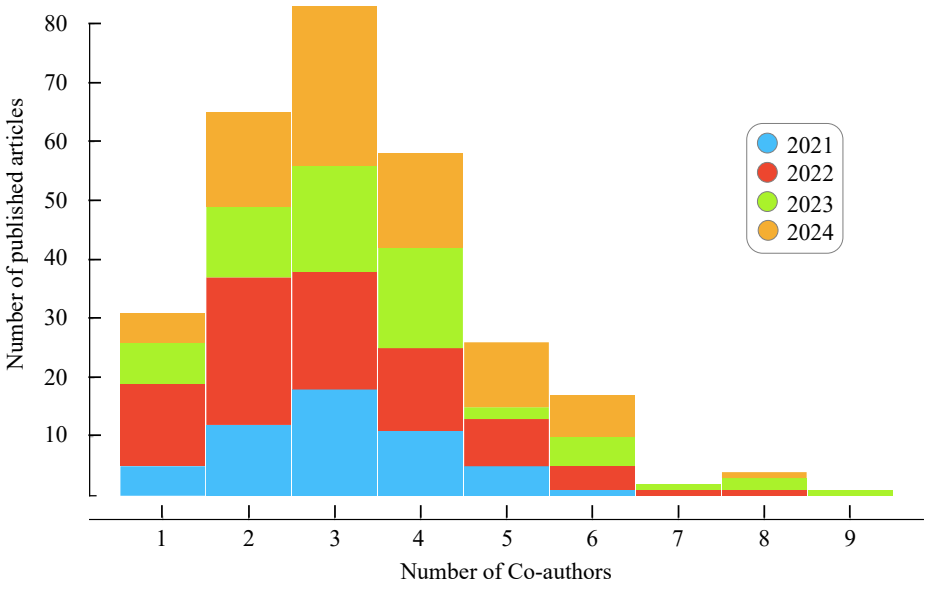


Figure 2. Variation in the number of co-authors across articles published by the journal.

Table 1. Number of published articles per year, and the variation in the number of co-authors.

year	number of co-authors									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**all**
2021	5	12	18	11	5	1	0	0	0	52
2022	14	25	20	14	8	4	1	1	0	87
2023	7	12	18	17	2	5	1	2	1	65
2024	5	16	27	16	11	7	0	1	0	83
Total	**31**	**65**	**83**	**58**	**26**	**17**	**2**	**4**	**1**	**287**

Note: n = 287 articles.

Figure 2 and Table 1 show that between 2021 and 2024, the journal received the most articles with 3 co-authors (83 articles), accounting for 29% of all published articles. From 2021 onward, the journal also began receiving articles with 7 co-authors (2 articles), 8 co-authors (4 articles), and up to 9 co-authors (1 article).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

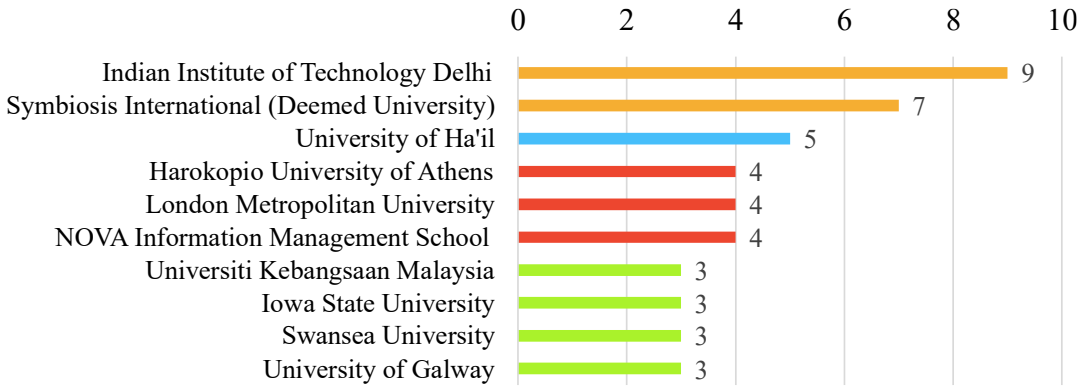


Figure 3. Educational institutions with the highest number of published articles in the journal, based on author affiliations.

Figure 3, based on the Scopus indexer (Scopus, 2024) shows that the Indian Institute of Technology Delhi was the educational institution with the most author affiliations that contributed to publications in the journal (9 times), followed by the institutions: Symbiosis International (Deemed University) - 7 times, University of Ha'il - 5 times, in addition to Harokopio University of Athens, London Metropolitan University, and NOVA Information Management School Lisboa - 4 times each. With a total of three published articles, we identified a larger list of educational institutions, and some of their names are shown in the figure.

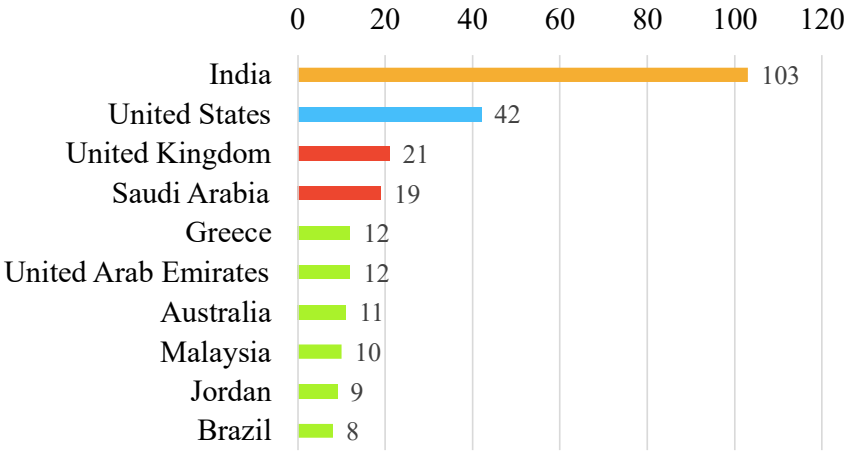


Figure 4. Number of documents published according to Country / Territory.

Figure 4 shows the list of Countries that are the headquarters of the Academic Institutions of the authors' affiliations that contributed most to publications in the journal, mainly India, with 103 articles published, followed by the United States of America (42), and then by the United Kingdom (21) and Saudi Arabia (19). Among the 12 to 8 publications in the journal, the countries are: Greece (12), United Arab Emirates (12), Australia (11), Malaysia (10), Jordan (9), and Brazil (8).

Table 2 identified, based on the variable *cita_tot*, the five articles published in the journal that received the highest number of citations – all of these were published in 2021, and two of them were cited more than 350 times, as assessed by the Scopus indexer (Scopus, 2024).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

Table 2. The five articles published in the journal that received the highest number of citations.

title	cit.	coa.	ref.	tit.	source
Fake news detection: A hybrid CNN-RNN based deep learning approach	357	3	63	66	Nasir et al. (2021)
Generative adversarial network: An overview of theory and applications	350	3	72	70	Aggarwal et al. (2021)
Artificial intelligence in marketing: Systematic review and future research direction	252	4	51	85	Verma et al. (2021)
Applications of big data in emerging management disciplines: A literature review using text mining	186	3	136	98	Kushwaha et al. (2021)
Applications of text mining in services management: A systematic literature review	186	3	144	82	(Kumar et al., 2021)

Notes: **cit**.=the number of citations from the Scopus journal indexer (Scopus, 2024); **coa**.=the number of co-authors; **ref**.=quantity of references; and **tit**.=number of characters existing in article titles including white spaces.

3.3. Cluster Analysis

To identify clusters of articles in the research sample (n=287), two clustering procedures were designed and implemented using the *Orange Data Mining* software. (Demšar et al., 2013; Godec et al., 2019): the first used the *Hierarchical agglomerative method* (S. Altuntas et al., 2022) and the second used the *K-means partitional method* (Husna et al., 2024).

Previously, to ensure rigor in the formation (modeling) of the clusters, the researchers verified the *Pearson correlation coefficient* among the four numerical variables chosen (*qty_ref, coaut_n, nc_tit, cita_tot*) and confirmed the absence of linear correlation and, even more, of a very low linear association among all of them, as shown in Figure 5.

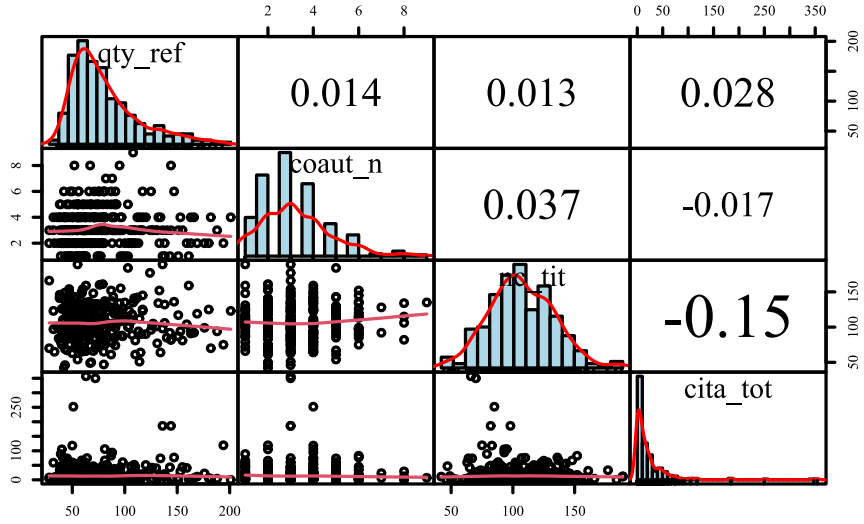


Figure 5. Pearson correlation matrix for the four variables used in clustering.

Notes: n = 287 articles published in the journal, between the years 2021 and 2024; *qty_ref* = quantity of references used in articles; *coaut_n* = number of co-authors who collaborated with the publication; *nc_tit* = number of characters used in the article title, including white spaces; *cita_tot* = total number of citations received by the article.

Pearson's correlation coefficient (Mu et al., 2018), also known as the product-moment correlation coefficient and usually represented by the lowercase letter '*r*', measures the strength of the linear

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

association between two variables, ranging from -1 to +1. The interpretation of *Pearson's linear correlation coefficient* may vary according to different areas of knowledge (Akoglu, 2018; Sadeghi, 2022; Schober et al., 2018). As the area of *Business, Accounting, and Management* is part of the broad knowledge area of *Applied Social Sciences*, the literature in this area (Cohen, 1988) reiterates that two variables have a strong positive linear correlation when the 'r' value is greater than or equal to +0.7, which does not hold in this case. As a result, the correlation matrix in Figure 5 indicates that correlations among the variables were very low or nonexistent; therefore, all variables could be used in cluster modeling.

When building multivariate quantitative models, the literature recommends performing multivariate outlier detection tests (Cabana et al., 2021; Leys et al., 2018) and the use of the *Mahalanobis distance* (Mahalanobis, 1930) is considered one of the most frequently indicated and applied tests.

Thus, the *Mahalanobis test* was applied to this set of four variables (*qty_ref, coaut_n, nc_tit, cita_tot*), with the use of the *R software* (R Core Team, 2026) and the *Performance package* (Lüdecke et al., 2021) and five multivariate outliers (which are particularly influential) were identified.

However, during the execution of quantitative modeling supported by statistical techniques, researchers are always responsible for making choices that are viable for the sequential conduct of analyses – and that is what occurred in this study. Because researchers already know the potential influence of outliers when clusters form (i.e., separations between groups of published articles), it was decided to retain these observations in the research dataset and carefully monitor the composition of the resulting clusters.

The first clustering technique that was carried out used the *Hierarchical agglomerative algorithm* (Lu et al., 2008) with the *Euclidean* distance metric and the *Ward* method for linkage. The *Euclidean* distance between two points (observations) is the length of the straight-line segment between them and can be calculated from their Cartesian coordinates using the Pythagorean theorem. Ward's linkage option was used to measure distances between clusters (Ward, 1963), which calculates the increase in the sum of squared errors, minimizing the total within-cluster variance (Murtagh & Legendre, 2014).

A *Dendrogram* is a representation of a tree of *hierarchical clusters* produced from a dataset, based on the available observations and on the choices made in the *distance measurement method* (e.g., Euclidean, Manhattan, Cosine, Jaccard, Pearson, etc.) and the *linkage method* (e.g., Single linkage, Average linkage, Weighted linkage, Complete linkage, Ward linkage). The graphical visualization of the *Dendrogram* (Murtagh & Legendre, 2014) in Figure 6 was generated using the *Hierarchical agglomerative clustering model*.

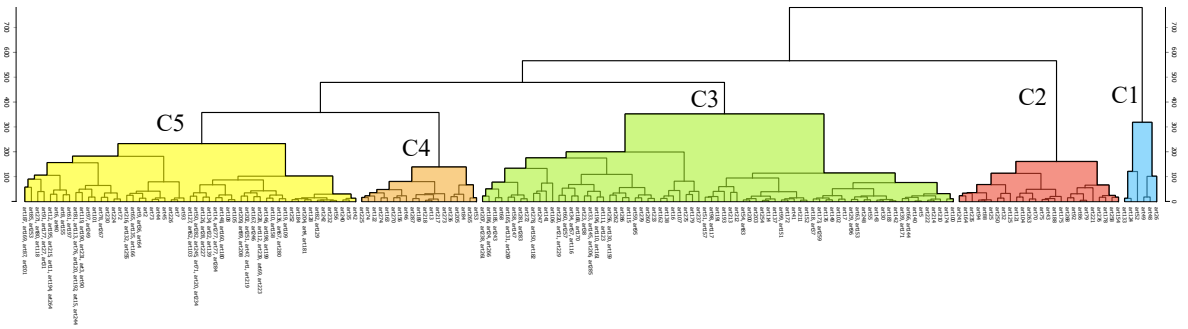


Figure 6. Dendrogram showing the composition of five clusters (C1, C2, C3, C4, and C5).

The second clustering technique employed a partitional K-means algorithm, and the optimal number of groups was determined using the Silhouette score. (Rousseeuw, 1987). As an alternative to using the *Elbow method,* the *Silhouette score* (Bagirov et al., 2023) is a method for interpreting and validating consistency within data clusters by comparing the mean distance to elements within the same cluster with the mean distance to elements in other clusters. It measures how similar an observation is to its own cluster (cohesion) relative to other clusters (separation).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

Using *Orange Data Mining software* (Demšar et al., 2013; Godec et al., 2019), it was possible to identify several *k* (being the number of clusters suggested by the *K-means algorithm*), and each of these *k* values has its respective *Silhouette score* value - the higher the *Silhouette score*, the better the clustering. In this case, the constitution of five clusters was the best choice, given a Silhouette score of 0.259 for *k=5*.

The *Silhouette score* is also calculated for each observation used in the clustering. A *Silhouette score* close to 1 indicates that the observation is close to the center of the cluster (the cluster is "better formed"), while observations with a Silhouette score close to 0 lie on the border between two clusters (more dispersed from the center of their respective clusters). A *Silhouette score* ranges from −1 to +1, where a high value indicates that the observation is well compatible with its own cluster and poorly compatible with neighboring clusters (Amorim & Hennig, 2015; Bagirov et al., 2023).

The K-means results also confirmed the number of clusters (k=5) generated by the Hierarchical agglomerative clustering algorithm and suggested changes to the positions of some measurements within each cluster. One of the algorithm's suggested changes was accepted, transferring one observation to cluster C2 (red; the observation located farthest up and to the right) and reducing the number of observations in cluster C1 from 6 to 5.

Table 3 and Figures 7, 8, and 9 show in detail how the five clusters (C1, C2, C3, C4, and C5) formed based on the four variables used in the two models.

Table 3. The four variables used to form the five clusters (C1, C2, C3, C4, and C5) and the types of measurements evaluated.

variable (label)	**measurement type**	**C1**	**C2**	**C3**	**C4**	**C5**
cluster name		Highly Cited	Heavily Documented	Descriptive Title	Mid Impact	Dominant Majority
quantity of references *(qty_ref)*	mean	93.2	152.6	79.7	54.1	62.9
	median	72	151	80	52	61
	Q1	63	132.5	55	46	51
	Q3	136	167.5	98.5	60.5	75
	min. value	51	110	28	32	28
	max. value	144	201	161	88	115
number of co-authors *(coaut_n)*	mean	3.2	3.1	3.3	2.7	3.2
	median	3	3	3	3	3
	Q1	3	2	2	1	2
	Q3	3	4	4	4	4
	min. value	3	1	1	1	1
	max. value	4	8	9	5	8
number of characters in title *(nc_tit)*	mean	80.2	94.4	130	108.2	89
	median	82	94	128	103	92
	Q1	70	83.5	119	93	75.8
	Q3	85	103	139.3	124	103
	min. value	66	64	89	79	42
	max. value	98	125	189	148	121
	mean	266.2	22.7	15	71	12.4

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

total	median	252	7	10	61	8
number of	Q1	186	2.5	1	55	1
citations(a)	Q3	350	41.5	22	82.5	18.3
(cita_tot)	min. value	186	0	0	45	0
	max. value	357	119	83	118	74
# in group	cluster					
	count (n)	5	27	112	19	124

Notes: n = 287 articles; the five clusters (C1, C2, C3, C4, and C5) were obtained using hierarchical clustering and the *k-means model*.

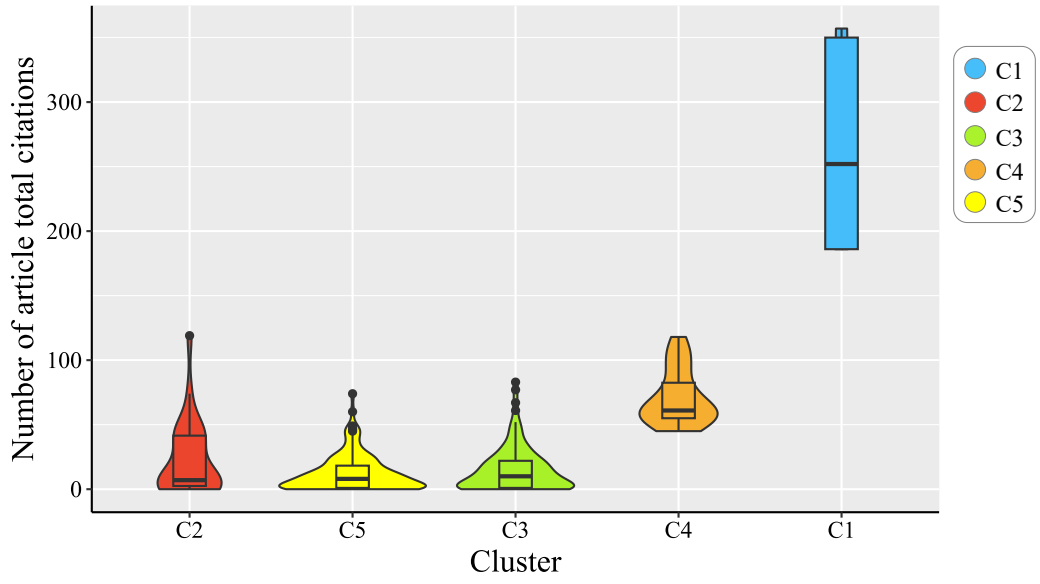
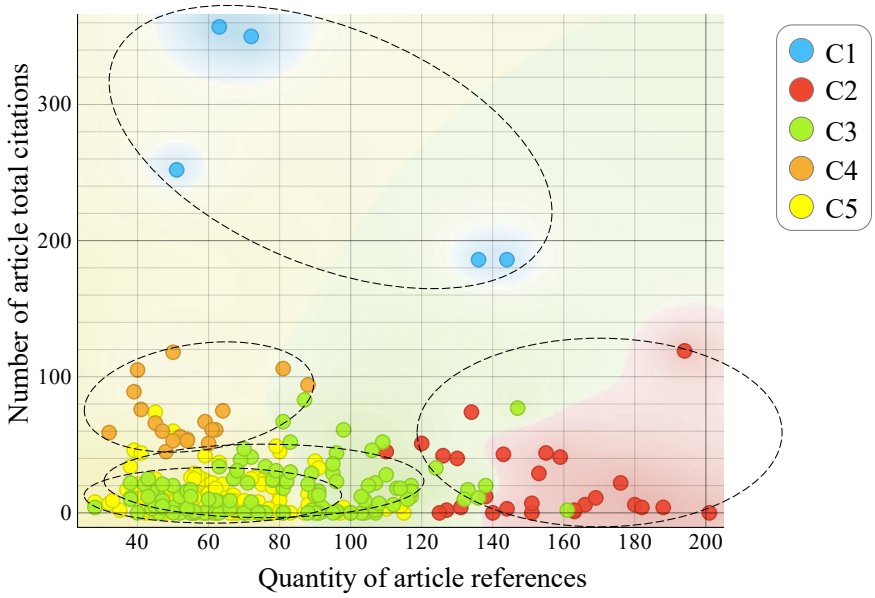


Figure 7. Violin Plot of the distribution of the number of total article citations (*cita_tot*) according to the five clusters (C1 – C5).

Notes: n = 287 articles; the measurements of each Box Plot (in the violin plot) are ordered by the median value; each of the five measurements (Mean, Median, Q1, Q3, Min. value, Max. value) per cluster is available in Table 3.



© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

Figure 8. Scatter Plot of the quantity of references (*qty_ref*) versus the total article citations (*cita_tot*), with the composition of five clusters (C1 – C5).

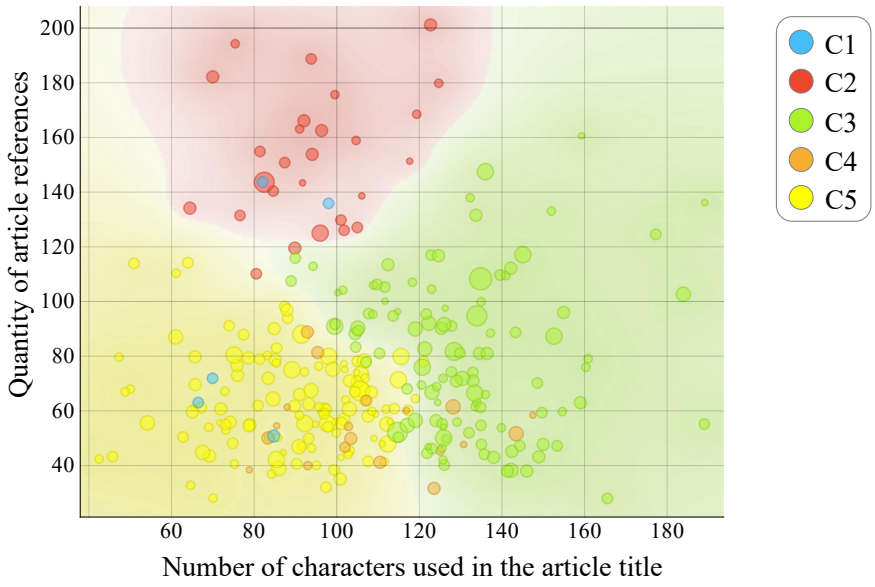


Figure 9. Scatter Plot of the number of characters used in the article title - including white spaces (*nc_tit*) versus the quantity of references (*qty_ref*), with the composition of five clusters (C1 – C5).

Based on Table 3 and Figures 7, 8, and 9, cluster C1 had the highest citation impact, with a median of 252 citations and an average of 266.2 citations, compared with 71 for the next. Although it had only 5 papers, these were considered particularly discrepant (outlier measurements) by the *Mahalanobis test* – see Table 2 (Aggarwal et al., 2021; Kumar et al., 2021; Kushwaha et al., 2021; Nasir et al., 2021; Verma et al., 2021), it appears to be the most influential group in the dataset, with well-established papers that have become foundational works in their field, a moderate number of references (mean=93.2), and relatively short titles (mean=80.2 characters). Because of this, *Cluster C1* was named "*Highly Cited Paper*".

Papers in cluster C2, composed of 27 observations, had a moderate title length (94.4 characters) and a median of 7 citations, the lowest among all clusters. This may indicate that they are recent, have undergone extensive review, or are still accruing citations. They had a high average of 152.6 references, nearly double that of the others, indicating extensive documentation or a broad study. *Cluster C2* was named "*Heavily Documented Research*".

Cluster C3, with 112 observations, had a moderate number of citations (median = 10; mean = 15) and the highest average number of characters in article titles (about 130), suggesting descriptive titles. As one of the larger groups with moderate references (mean = 79.7), it likely reflects detailed research or standard empirical papers with specific scopes. *Cluster C3* was named "*Descriptive Title Research*".

Cluster C4, composed of 19 observations, was in the opposite direction of cluster C2, considered the group with the lowest median number of references used in the articles (around 52 references), and was like cluster C3 in being the second cluster with a high number of characters in article titles (mean of 108.2 characters). However, the second group had the highest number of citations (median of 61, mean of 71) – see Figure 9. These appear to be targeted empirical studies with a narrower scope but strong citation impact. *Cluster C4* was named "*Mid Impact Study*".

In data analysis, the group with the largest number of members generally defines what is considered the "typical/standard" or most common behavior in that specific sample or population studied.

In this research, cluster C5, the largest group (124 observations), represented the "norm" (or standard) because it was the most representative in terms of statistics and frequency (43% of the sample, n=287), describing the typical situation in which most articles are found. Unlike clusters C1

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

(with extreme citation counts) or C2 (with a very high number of references), cluster C5 had moderate values (median of 61 references, average of 89 characters in the title) and lower values (average of 12.4 citations), suggesting that the vast majority of articles in this dataset have a more modest citation impact rather than being exceptional cases. *Cluster C5* was named "*Dominant Majority Publication*".

3.4. Text Mining

Initially, from all the titles of the 287 articles that made up the data sample, each *unigram* token was extracted, previously excluding all *stopwords* (Favaretto, 2024a). *Bigrams* and *trigrams* were also extracted, obtaining the total number of occurrences (frequency) of each, as well as the specific position (rank) they occupied, according to their type (unigram, bigram, or trigram).

Then, this list, comprising the integration of the three token types, was manually checked (visually) to confirm whether the extracted tokens made semantic sense in the contexts addressed in the set of publications. At the end of this procedure, 189 tokens (*n-grams*) were identified, including 125 *unigrams* with frequencies of up to five occurrences, shown in the word cloud in Figure 10 and in the list presented in Table 4 (partially).

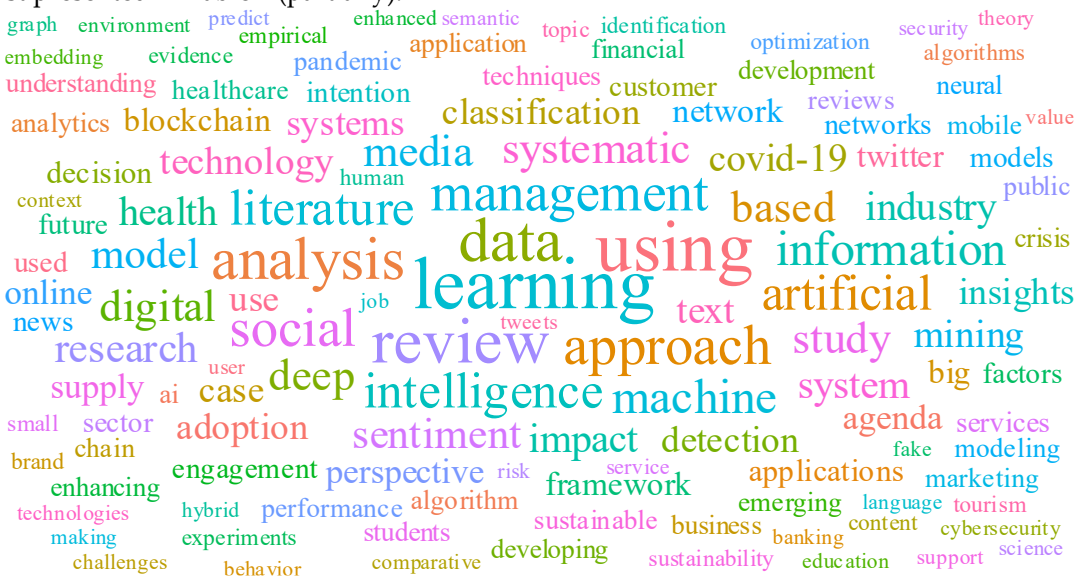


Figure 10. Word cloud of *unigrams*, with a frequency of up to five occurrences, found in article titles.

Table 4. Sorted list of the first 51 unigrams, ranked by decreasing frequency of occurrence (freq.) and increasing position (rank), extracted from the 189 tokens identified in the titles of the articles in this research.

unigram	freq.	rank	unigram	freq.	rank	unigram	freq.	rank
learning	54	1	systematic	19	18	use	14	31
using	52	2	media	19	18	systems	14	31
review	41	3	model	18	20	insights	13	37
data	39	4	research	18	20	supply	12	38
analysis	35	5	sentiment	18	20	perspective	12	38
approach	32	6	system	16	23	agenda	12	38
management	29	7	impact	16	23	online	12	38
social	28	8	health	16	23	twitter	11	42
literature	26	9	industry	16	23	framework	11	42
intelligence	26	9	covid-19	15	27	big	11	42
artificial	24	11	mining	15	27	network	11	42
information	23	12	text	15	27	blockchain	11	42

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

machine	23	12	technology	15	27	decision	10	47
study	22	14	case	14	31	applications	10	47
based	22	14	detection	14	31	engagement	9	49
deep	21	16	classification	14	31	networks	9	49
digital	21	16	adoption	14	31	future	9	49

Note: The rank position repeats when the *n-gram* has the same occurrence count (freq.) as the previous one. When the rank position changes, it moves to the position immediately after the absolute ordering that was in progress.

.

Also, a total of 49 *bigrams* and 15 *trigrams* with a frequency of up to three occurrences were identified, as shown in Figure 11 and Table 5.

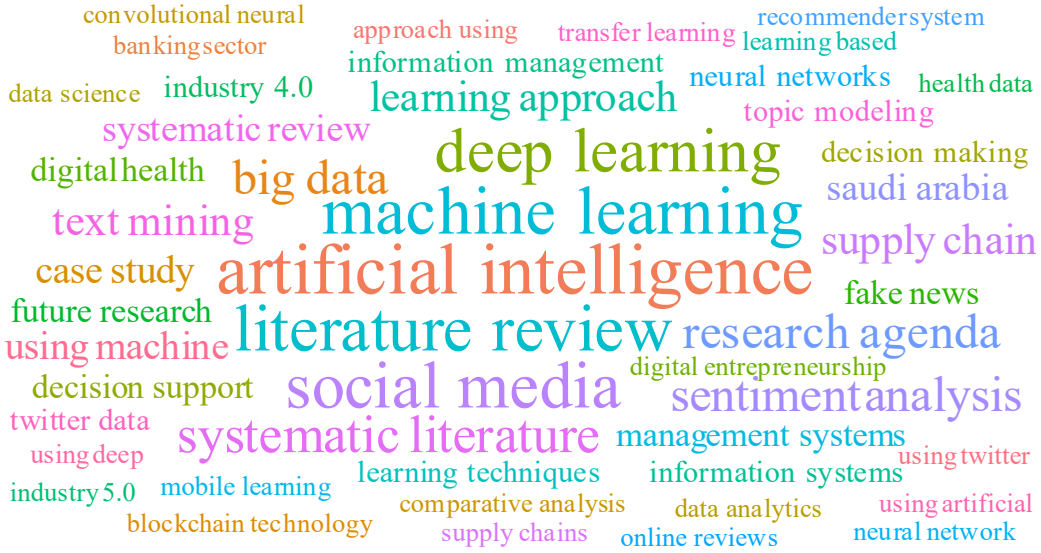


Figure 11. Word cloud of *bigrams*, with a frequency of up to three occurrences, found in article titles.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

Table 5. Sorted list of the first 34 *bigrams*, ranked by decreasing frequency of occurrence (freq.) and increasing position (rank), extracted from the 189 tokens identified in the titles of the articles in this research.

bigram	freq.	rank	bigram	freq.	rank
artificial intelligence	22	1	management systems	5	17
machine learning	20	2	decision support	5	17
literature review	19	3	fake news	5	17
social media	19	3	digital health	5	17
deep learning	17	5	decision making	4	22
systematic literature	12	6	topic modeling	4	22
research agenda	11	7	learning techniques	4	22
big data	11	7	information management	4	22
sentiment analysis	10	9	information systems	4	22
text mining	9	10	industry 4.0	4	22
learning approach	8	11	neural networks	4	22
supply chain	8	11	twitter data	4	22
case study	7	13	mobile learning	3	31
using machine	7	13	using deep	3	31
systematic review	6	15	transfer learning	3	31
saudi arabia	6	15	data analytics	3	31
future research	5	17	data science	3	31

Note: The rank position repeats when the *n-gram* has the same occurrence count (freq.) as the previous one. When the rank position changes, it moves to the position immediately after the absolute ordering that was in progress.

To visually identify 18 trigrams and their respective frequencies of occurrence in the titles of articles published in the journal, Figure 12 and Table 6 were generated.

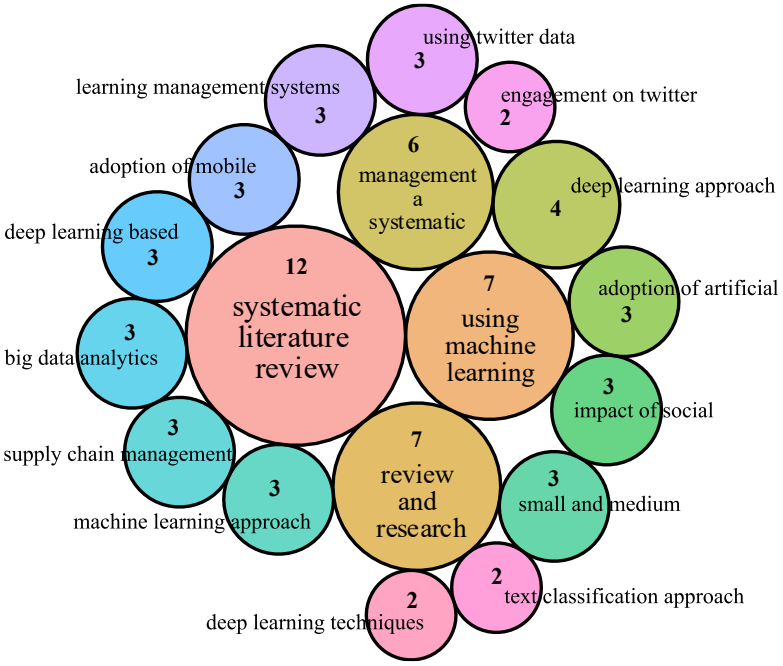


Figure 12. Circle chart of *trigrams*, with a frequency of up to two occurrences, identified in article titles.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

Table 6. Sorted list of the first 18 *trigrams*, ranked by decreasing frequency of occurrence (freq.) and increasing position (rank), extracted from the 189 tokens identified in the titles of the articles in this research.

trigram	freq.	rank	trigram	freq.	rank
systematic literature review	12	1	supply chain management	3	6
using machine learning	7	2	big data analytics	3	6
review and research	7	2	deep learning based	3	6
management a systematic	6	4	adoption of mobile	3	6
deep learning approach	4	5	learning management systems	3	6
adoption of artificial	3	6	using twitter data	3	6
impact of social	3	6	engagement on twitter	2	16
small and medium	3	6	text classification approach	2	16
machine learning approach	3	6	deep learning techniques	2	16

Note: The rank position repeats when the *n-gram* has the same occurrence count (freq.) as the previous one. When the rank position changes, it moves to the position immediately after the absolute ordering that was in progress.

3.5. Social Network Analysis

Words *co-occur* when two or more words appear in a block of text selected for this analysis. For example, if this analysis is performed on several article titles, the collection of article titles must be *stacked* in a single column, with each line representing a different title. The analysis will be done per line and will consider the *co-occurrence* of words in all possible combinations of two words that appear in that line, not limited to those that are positioned together (adjacent). By the example given, it can be said that two words *co-occur* when they are mentioned at least once in the same article title, regardless of their position (i.e., they co-occur). The more article titles contain these two words, the stronger (higher frequency of co-occurrence) the connection (link, tie, or edge) between them will be.

From the 189 n-grams extracted from the titles of the articles – unigram (125 simple terms, 66.2% of the network), bigram (49 expressions with two terms, 25.9% of the network), and trigram (15 expressions with three terms, 7.9% of the network) – now considered nodes of a network, direct co-occurrences between these terms and expressions were generated, resulting in a total of 6912 edges (directed graph among these nodes).

To allow exploration of SNA techniques (F. Altuntas et al., 2022) in this portion of the research dataset, Gephi software was used (Bastian et al., 2009) starting by converting edges from a direct graph to an indirect graph, which is more suitable for analyzing word co-occurrences; the total number of edges in this network was reduced from 6912 to 4014.

A necessary and essential step in generating a network graph appropriate to the research interest is the choice of a *layout algorithm* from among several existing ones, based on the topology feature you want to highlight (Gephi, 2011): emphasis divisions - OpenOrd,

emphasis complementarities – ForceAtlas, Yifan Hu (Hu, 2006), Fruchterman-Reingold (Fruchterman & Reingold, 1991), emphasis ranking – Circular, Radial Axis, and emphasis geographic repartition – GeoLayout.

Basically, *layout algorithms* define the graph's shape by determining node positions within the graph panel, following a simple *force-based* principle: *linked nodes* attract each other, and *unlinked nodes* repel each other.

The network in Figure 13 was formed only by bigrams and trigrams at its nodes – all with a minimum of 3 occurrences – and using the Fruchterman-Reingold layout model (Fruchterman & Reingold, 1991), which ensures that nodes in a network minimize crowding or even overlap, are distributed more evenly throughout the network, and facilitate the identification of possible groupings. The sizes of the nodes were generated using the eigen centrality measure, and the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

variation in colors (nodes and edges) was generated using the community detection method in the *modularity class* function (Blondel et al., 2008).

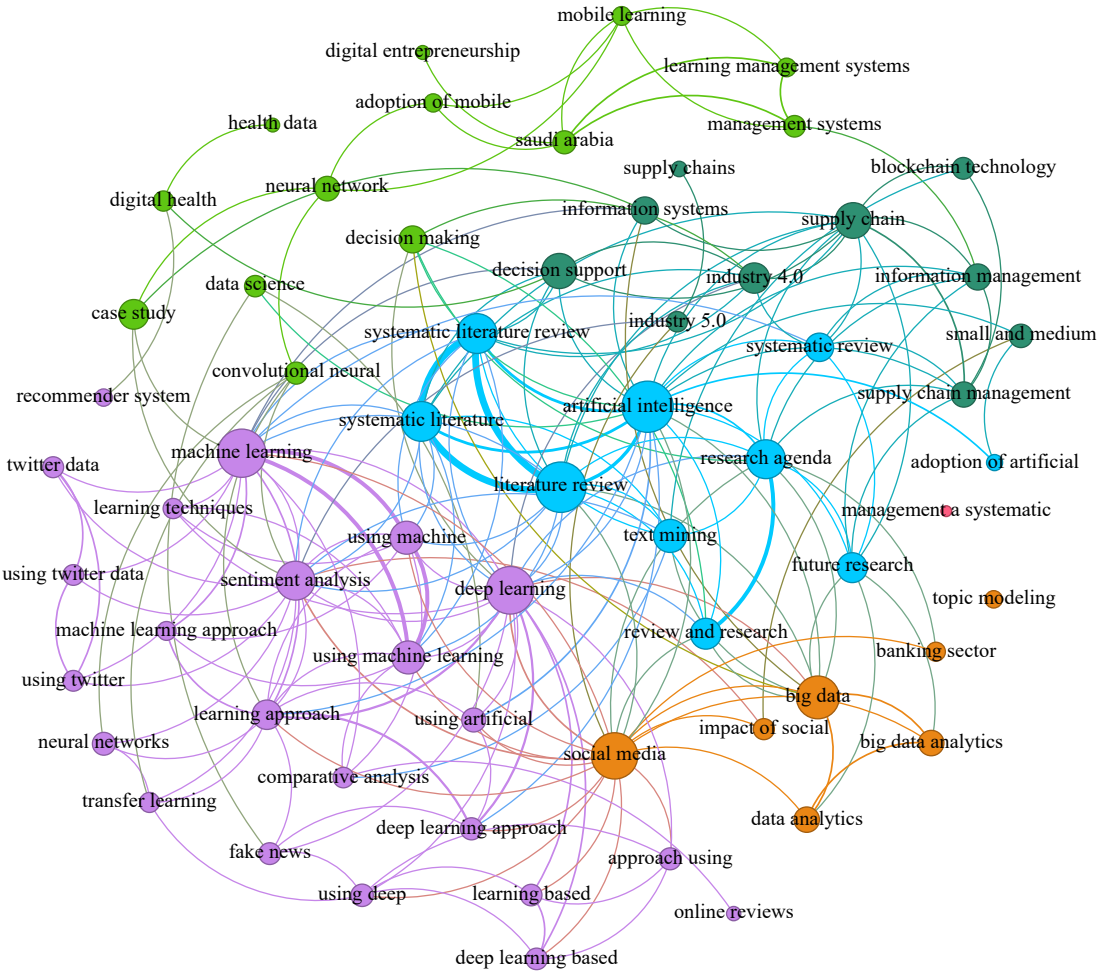


Figure 13. Network of word *co-occurrences* identified in article titles, composed of *bigrams* and *trigrams*, with higher eigen centralities (node sizes) and different modularity classes (node and edge colors).

Table 7. Network measures of word *co-occurrences*.

n-gram	freq.	rank	group	eigen.	mod.	degree	betweenness
artificial intelligence	22	1	bigram	0.7109	1	84	171.2306
literature review	19	3	bigram	0.6968	1	73	69.0664
machine learning	20	2	bigram	0.6585	3	78	154.9218
deep learning	17	5	bigram	0.6466	3	75	135.0870
social media	19	3	bigram	0.6243	2	76	167.6063
big data	11	7	bigram	0.5634	2	60	52.9715
systematic literature	12	6	bigram	0.5103	1	53	32.0362
systematic literature review	12	1	trigram	0.5103	1	53	32.0362
sentiment analysis	10	9	bigram	0.4939	3	57	75.1452
research agenda	11	7	bigram	0.4937	1	52	35.3016
supply chain	8	11	bigram	0.4375	5	46	20.2780
decision support	5	17	bigram	0.4277	5	43	21.9837
text mining	9	10	bigram	0.3922	1	42	31.6678

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

using machine	7	13	bigram	0.3903	3	40	18.9373
using machine learning	7	2	trigram	0.3903	3	40	18.9373
review and research	7	2	trigram	0.3531	1	33	5.6784
industry 4.0	4	22	bigram	0.3435	5	36	14.0348
future research	5	17	bigram	0.3360	1	36	13.8986
case study	7	13	bigram	0.3320	0	36	17.9055
learning approach	8	11	bigram	0.3276	3	38	28.5248
systematic review	6	15	bigram	0.3097	1	33	11.2199
information systems	4	22	bigram	0.2866	5	27	5.3631
decision making	4	22	bigram	0.2850	0	30	11.0895
information management	4	22	bigram	0.2693	5	29	7.6756
data analytics	3	31	bigram	0.2567	2	27	5.0219
big data analytics	3	6	trigram	0.2567	2	27	5.0219
supply chain management	3	6	trigram	0.2539	5	27	5.2217
neural network	3	31	bigram	0.2432	0	31	16.8373
using artificial	3	31	bigram	0.2310	3	22	3.4460
small and medium	3	6	trigram	0.2219	5	25	15.1257
saudi arabia	6	15	bigram	0.2150	0	28	21.4363
neural networks	4	22	bigram	0.2118	3	22	5.7196
using deep	3	31	bigram	0.2038	3	24	7.9143
deep learning approach	4	5	trigram	0.1990	3	20	2.9905
convolutional neural	3	31	bigram	0.1978	0	21	4.0087
approach using	3	31	bigram	0.1963	3	20	3.2612
fake news	5	17	bigram	0.1930	3	24	8.9697
blockchain technology	3	31	bigram	0.1923	5	21	5.3254
learning based	3	31	bigram	0.1896	3	19	2.1006
deep learning based	3	6	trigram	0.1896	3	19	2.1006
management systems	5	17	bigram	0.1889	0	23	8.6703
data science	3	31	bigram	0.1862	0	21	6.8968
impact of social	3	6	trigram	0.1805	2	20	4.7645
twitter data	4	22	bigram	0.1785	3	18	1.8463
using twitter	3	31	bigram	0.1749	3	17	0.9879
using twitter data	3	6	trigram	0.1749	3	17	0.9879
comparative analysis	3	31	bigram	0.1713	3	18	5.7574
industry 5.0	3	31	bigram	0.1687	5	17	5.6269
digital health	5	17	bigram	0.1656	0	21	11.5854
learning techniques	4	22	bigram	0.1652	3	16	0.8722
transfer learning	3	31	bigram	0.1621	3	18	2.7275
banking sector	3	31	bigram	0.1438	2	17	2.0166
machine learning approach	3	6	trigram	0.1418	3	14	0.9624
mobile learning	3	31	bigram	0.1407	0	21	6.5539

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

learning management systems	3	6	trigram	0.1382	0	17	1.4943
adoption of mobile	3	6	trigram	0.1310	0	19	4.5202
topic modeling	4	22	bigram	0.1159	2	14	1.0085
recommender system	3	31	bigram	0.1146	3	14	3.2565
adoption of artificial	3	6	trigram	0.0938	1	10	0.4446
supply chains	3	31	bigram	0.0825	5	10	0.7160
online reviews	3	31	bigram	0.0612	3	8	0.7761
digital entrepreneurship	3	31	bigram	0.0554	0	10	1.4203
health data	3	31	bigram	0.0530	0	6	0.0000
management a systematic	6	4	trigram	0.0000	4	0	0.0000

Note: The table is ordered by the eigen centralities of **n-grams** (bigrams and trigrams) extracted from the article titles; **freq.** = word frequency (count); **rank** = ranking position by n-gram type; **group** = bigram or trigram; **eigen**. = eigenvector centrality (measures a node's importance based on the importance of its neighbors); **mod**. = indicates the *modularity class* and was generated by the *Louvain method*; **degree** = degree centrality (measures the number of direct connections a node has, indicating an element's popularity or immediate activity); **betweenness** = betweenness centrality (measures how often a node sits on the shortest paths between other pairs of nodes).

Modularity measures the strength of dividing a network graph into groups, or the quality of partitioning a network into clusters. The Gephi software uses the *Louvain method* (Blondel et al., 2008, 2024) to extract *modularity classes* and thus discover communities among network nodes. After running the modularity algorithm, six node groups (communities composed of bigrams and trigrams) were identified, numbered 0 to 5 (as shown in the "mod." column in Table 7), and each node in the network belonged to one of these groups.

In Figure 13 and Table 7, *grouping 0* was identified by the light green color, consisted of 13 nodes (20.1% of the network), and included the *n-grams*: mobile learning, adoption of mobile, learning management systems, digital entrepreneurship, health data, digital health, decision making, data science, neural network, convolutional neural, case study, and Saudi Arabia. Most *n-grams* in this cluster had an eigen centrality value of 0.2432 or lower and a frequency of 3, focusing on themes of mobile, health, decision-making, and the application of neural network techniques.

Identified by its light blue color and composed of 10 nodes (17.9% of the network), *grouping 1* included the *n-grams*: artificial intelligence, adoption of artificial, text mining, systematic literature review, review and research, research agenda, and future research. This grouping was characterized by central discussions in the journal, demonstrating broad interest in artificial intelligence, methodological aspects of literature reviews, and the positioning of the current and future research agenda.

Grouping 2, in orange, comprised 7 nodes (16.9% of the network) and included the *n-grams*: social media, the

impact of social, big data, data analytics, topic modeling, and the banking sector. It highlighted the authors' broad interest in social media and big data, which were mentioned 19 and 11 times, respectively, in the journal's articles.

As shown in Figure 13 (purple), *grouping 3* was formed by 23 nodes (28.6% of the network) – the largest number of nodes among all other groupings – and included the *n-grams*: using Twitter data, online reviews, fake news, sentiment analysis, recommender system, comparative analysis, beyond n-grams: using artificial, neural networks, using machine learning, deep learning, deep learning based, deep learning approach, learning techniques, and transfer learning. This grouping highlighted two other central themes in the journal, both associated with artificial intelligence: machine learning (20 occurrences, eigen centrality of 0.6585, betweenness centrality of 154.9218) and deep learning (17

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

occurrences, eigen centrality of 0.6466, betweenness centrality of 135.0870). *Grouping 4*, formed by a single node (red, 0.5% of the network), in isolation, produced the *trigram* 'management a systematic', reflecting the focus of the publications identified in the journal, with an emphasis on systematic literature reviews.

Finally, *grouping 5*, in dark green, made up of 10 nodes (16% of the network), identified the n-grams: supply chain(s) [management], decision support, industry 4.0 and 5.0, information systems, information management, small and medium, and blockchain technology, highlighting emerging themes in business & management, associated with technology, and addressed by the journal's publications.

3.6. Topic Modeling

To perform *Topic Modeling* analysis using a set of titles, keywords, and abstracts from 287 scientific texts, and the R software, it was necessary to follow a series of steps that included data preprocessing, text cleaning, feature extraction, model selection, and results evaluation, i.e.: i) Data Import: load the data from the titles, keywords, and abstracts of 287 articles; ii) Text Preprocessing: combine these 3 full content items into a single text column, clean the text data (e.g., remove stop words, punctuation, and non-alphanumeric characters), tokenize the text, and apply stemming to words (the '*wordStem*' function in R performs word stemming, that is, it reduces each word to its root/radical, removing suffixes and endings - for example: "studies", "studying", "studied" become "studi"); iii) Vectorization and Model Selection: convert the text data (terms/words) into a numerical representation (vectors) using count vectorization, also known as Bag of Words (BoW), because the Latent Dirichlet Allocation (LDA), the topic modeling algorithm used in this work, doesn't understand words, it operates on numbers and requires raw counts, not normalized weights; iv) Model Training: fit the LDA model to the vectorized data; and v) Results: the topics were extracted from the model, and the description and visualization of these results were presented through the generation of Figures 14 and 15, and Table 8.

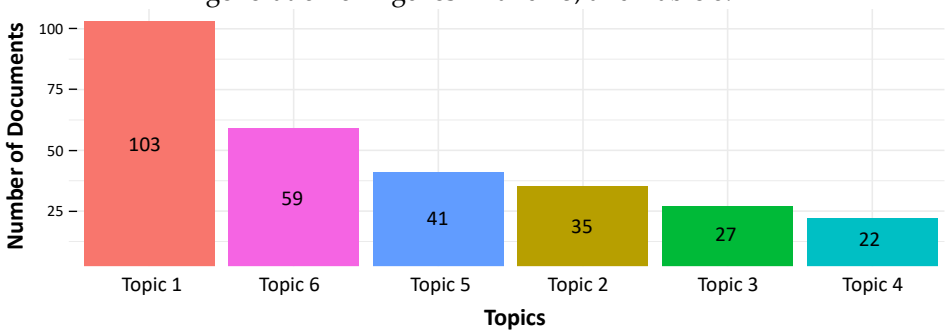


Figure 14. Number of documents that make up each of the six topics

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

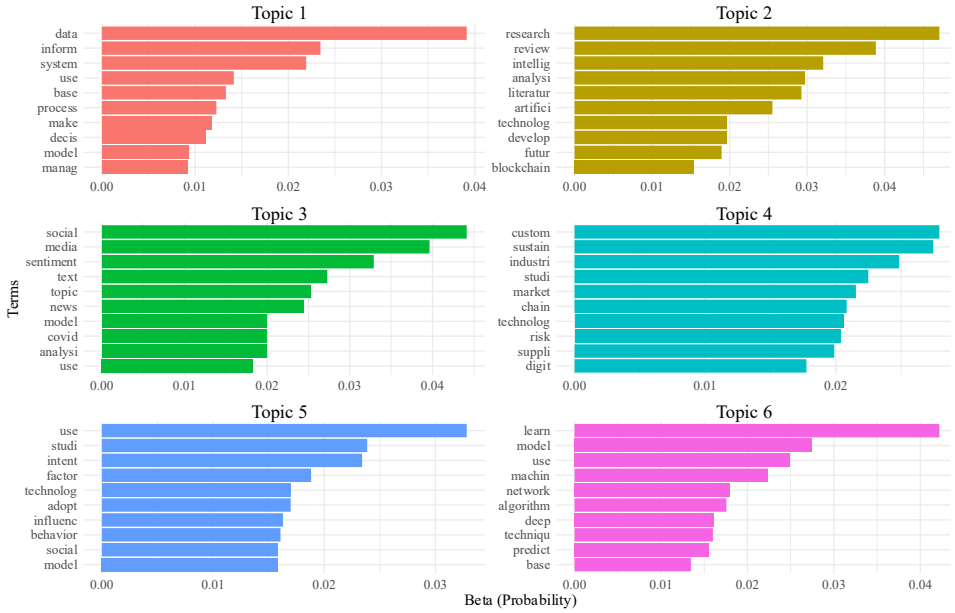


Figure 15. Composition of the six topics with identification of their words and probabilities.

Table 8. Words identified and the probability of each word belonging to each of the six topics.

word	probability	topic	word	probability	topic	word	probability	topic
data	0.039136	1	research	0.047057	2	social	0.044133	3
inform	0.023448	1	review	0.038863	2	media	0.039636	3
system	0.021913	1	intellig	0.032074	2	sentiment	0.032890	3
use	0.014154	1	analysi	0.029733	2	text	0.027268	3
base	0.013302	1	literatur	0.029265	2	topic	0.025301	3
process	0.012279	1	artifici	0.025520	2	news	0.024457	3
make	0.011767	1	technolog	0.019667	2	covid	0.019960	3
decis	0.011085	1	develop	0.019667	2	analysi	0.019960	3
model	0.009294	1	futur	0.018965	2	model	0.019960	3
manag	0.009209	1	blockchain	0.015453	2	use	0.018273	3
custom	0.027934	4	use	0.032811	5	learn	0.042213	6
sustain	0.027460	4	studi	0.023863	5	model	0.027482	6
industri	0.024857	4	intent	0.023404	5	use	0.024932	6
studi	0.022490	4	factor	0.018816	5	machin	0.022382	6
market	0.021543	4	technolog	0.016980	5	network	0.017991	6
chain	0.020833	4	adopt	0.016980	5	algorithm	0.017566	6
technolog	0.020596	4	influenc	0.016292	5	deep	0.016150	6
risk	0.020359	4	behavior	0.016062	5	techniqu	0.016008	6
suppli	0.019886	4	social	0.015833	5	predict	0.015583	6
digit	0.017755	4	model	0.015833	5	base	0.013458	6

Based on analyses of the extracted terms (stems), their respective probabilities (Beta), and the researcher's direct queries of the articles in the dataset, the six topics identified in the sample of 287 scientific articles could be interpreted appropriately.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

Topic 1: Data Management and Decision Support (103 articles, 35.8% of the sample studied) - this topic, based on the words "data", "inform", "system", "use", "base", "process", "make", "decis", "model" and "manag" focuses on information infrastructure, management information systems, and data-driven decision-making processes. This topic covers various aspects of information systems and data management to support and optimize organizational processes, information retrieval, and decision-making. The article focuses on process mining and data analysis to improve compliance, efficiency, and quality in organizational systems, including analyzing business workflows, verifying compliance in software development, and conducting electronic internal audits.

Topic 2: Artificial Intelligence and Emerging Technological Frontiers (35 articles, 12.2% of the sample) - the focus of this topic is academic and forward-looking. The presence of "intellig", "artifici", and "blockchain" indicates research focused on literature reviews on emerging technologies, especially blockchain and Artificial Intelligence (AI). Other words also belong to this topic: "research", "review", "intellig", "analysi", "literatur", "artifici", "technolog", "develop", "futur", "blockchain". In this topic, the papers focus on systematic literature reviews and bibliometric analyses to map the state of the art and future trajectories of emerging digital technologies and their applications across multiple domains. The works use network mapping and text mining to synthesize existing knowledge and propose future research agendas. The technologies examined include artificial intelligence, blockchain, big data, virtual and augmented reality, and cryptocurrencies, applied in contexts such as digital marketing, virtual tourism, the metaverse, and cybersecurity.

Topic 3: Social Media Analysis and Sentiment Mining (27 articles, 9.4% of the sample) - this topic brought together articles clearly focused on analyzing social media text data. In the word list of this topic ("model", "social", "media", "sentiment", "text", "topic", "news", "model", "covid", "analysis", “use”), the term "covid" suggests that part of the sample of articles analyzed behavior or news in the context of the COVID-19 pandemic, through text mining and/or sentiment analysis. The documents in this topic analyze unstructured social media data, particularly Twitter data, to understand human behavior, social dynamics, and crisis communication. The central focus of the research is the application of Natural Language Processing (NLP) techniques, sentiment analysis, and topic modeling to investigate public opinion and engagement across contexts, monitor mental health, predict content popularity, and identify the spread of fake news, particularly in critical contexts such as the COVID-19 pandemic and natural disasters.

Topic 4: Supply Chain and Sustainability (22 articles, 7.7% of the sample) - the terms in this topic ("custom", "sustain", "industri", "market", "chain", "study", "technolog", "risk", "suppli", "digit") point to studies on sustainability, risks and digitalization, digital transformation, and sustainable management in supply chains. This topic examines the convergence of sustainability and digital transformation in supply chain management and consumer behavior. It addresses the intersection of digital technologies, sustainability, and organizational transformation across sectors, including retail, e-commerce, supply chain, hospitality, banking, construction, and financial services. A recurring theme is how customer expectations, environmental awareness, and sustainability are transforming traditional business models. The central focus is on transitioning to a green, service-oriented economy, where Industry 4.0 technologies - such as traceability, process digitization, and logistics tools - are essential to strengthening organizational resilience and environmental compliance.

Topic 5: Technological Adoption and Behavior (41 articles, 14.3% of the sample) - this topic explores the factors that influence the intention to use/adopt new technologies. It focuses on user behavior, human behavior, and social factors related to adopting technology and/or innovation. The words "use", "studi", "intent", "factor", "technolog", "adopt", "influenc", "behavior", "social" and "model" were highlighted in this topic. This topic brings together documents that analyze the psychological, social, and technological determinants that influence intention to adopt and behavior in using digital innovations. Using established theoretical frameworks for behavior and technology acceptance, such as the Theory of Planned Behavior (TPB) and the Unified Theory of Acceptance and Use of Technology (UTAUT), they address themes including technology adoption and acceptance, behavioral intention, and the factors that influence the use of digital systems across contexts.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

Topic 6: Machine/Deep Learning and Predictive Modeling (59 articles, 20.6% of the sample) - this topic focuses on the technical and algorithmic aspects. The terms "learn", "machin", "deep", and "predict" directly refer to machine learning and deep learning techniques for prediction tasks (predictive models and algorithms). Other words also present in this topic are: "model", "use", "network", "algorithm", "techniqu" and "base." The articles in this topic address the development and diverse applications of deep learning and machine learning techniques to solve complex classification and prediction problems across multiple technological domains. The works employ hybrid or advanced neural network architectures, including convolutional, recurrent, adversarial, and transformer models, to address problems ranging from human activity recognition and deepfake detection to sales forecasting and financial market analysis.

Table 9 presents, in an integrated format, the six dominant topics explored in the research, including the number of articles per topic (column n), the words associated with each topic, and the assigned name for each topic.

Table 9. The six dominant topics explored in the research.

topic	n	words	broader topic
1	103	data, inform, system, use, base, process, make, decis, model, manag	Data Management and Decision Support
2	35	research, review, intellig, analysi, literatur, artifici, technolog, develop, futur, blockchain	Artificial Intelligence and Emerging Technological Frontiers
3	27	social, media, sentiment, text, topic, news, covid, analysi, model, use	Social Media Analysis and Sentiment Mining
4	22	custom, sustain, industri, studi, market, chain, technolog, risk, suppli, digit	Supply Chain and Sustainability
5	41	use, studi, intent, factor, technolog, adopt, influenc, behavior, social, model	Technological Adoption and Behavior
6	59	learn, model, use, machin, network, algorithm, deep, techniqu, predict, base	Machine/Deep Learning and Predictive Modeling

Table 9 shows that Topic 1 (Data Management and Decision Support) emphasizes broader cross-cutting themes across multiple research types, whereas Topic 6 (Machine/Deep Learning and Predictive Modeling) emphasizes more specialized niches.

Another way to understand a document's influence on a specific topic is to examine the data visualization in the form of a heatmap. The heatmap shown in Figure 16 was generated from a stratified sample of 54 documents, selecting the 9 documents with the highest probabilities (most representative) for each of the 6 topics to ensure balanced coverage in this visualization. Table 10 also interprets the heatmap's color scale, indicating the probability that a specific document is more closely associated with a particular topic.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

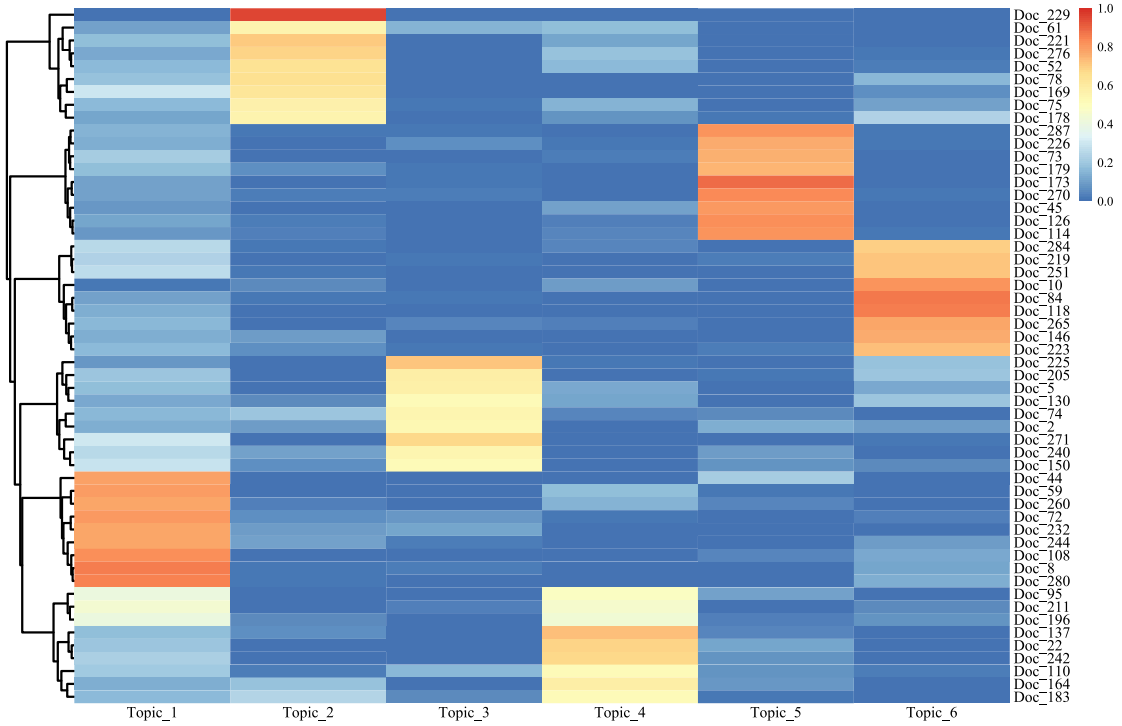


Figure 16. Heatmap of the nine documents with the highest probabilities for each topic.

Table 10. Interpretations of the heatmap color scale markings.

heatmap scale	color	probability level	meaning
0.8 - 1.0	red (hot)	high	The document strongly belongs to this topic
0.6 - 0.8	orange	moderate-high	
0.4 - 0.6	yellow	moderate	
0.2 - 0.4	light blue	low	
0.0 - 0.2	dark blue (cold)	very low	The document barely belongs to this topic

To interpret the heatmap in Figure 16 and Table 10, note that each row represents one of the 54 selected documents, each column identifies a topic (1 to 6), and each cell's color indicates the probability that the document belongs to that topic. Red cells (hot) indicate a strong association (high probability), and dark blue cells (cold) indicate a weak association (very low probability). For example, based on the data generated in this research, it can be observed that the document Doc_84 (Gupta, 2021) - an article that contributed to the discussion on recognizing human activities through deep learning with data from wearable sensors, proposing a hybrid model - has a red cell (probability of 0.862689) in Topic 6, meaning there is an 86.3% probability that this document belongs to Topic 6 (Machine/Deep Learning and Predictive Modeling).

As a second example, when considering the document Doc_108 (Kouzari et al., 2023) - which investigated the use of process mining for verifying process conformance in the development of enterprise information systems - also has a red cell (probability of 0.825508) in Topic 1, meaning there is an 82.6% probability that this document belongs to Topic 1 (Data Management and Decision Support).

It is also noticeable that light blue or dark blue markings appear in all other columns for the other topics, indicating a low or very low probability that the exemplified documents belong to those topics.

Another point to note when interpreting the heatmap is the hierarchical agglomerative clustering (visualized by the dendrogram on the left side of Figure 16), which used Euclidean distance

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

and the complete method as parameters and grouped nearby document branches with similar topic distributions.

4. Discussion

Actionable understanding from datasets on information management and data insights should improve researchers' intellectual knowledge of the field of Data, Information, and Management, whether through the selection of analytical techniques and algorithmic modeling applications, or through the interpretation of results obtained from these techniques/models.

This work has shown and discussed the general difference between DM and ML: while DM focuses on data in all forms as an application domain and aims to extract patterns, ML is a technology centered on methods that enable computers to learn (Coenen, 2011). In this way, social network analysis, text mining, and co-occurrence analysis could be considered analytical approaches that align with DM, as they use algorithms to discover hidden patterns in unstructured or relational data, such as social interaction networks, word frequencies in texts, and associations between terms, resulting in the extraction and discovery of non-trivial knowledge that can be used to identify new opportunities for conducting research in the studied field. Clustering and LDA could be considered ML models.

Because the intuitions behind the LDA model assign a dominant topic to each article based on relevance and group related articles by theme, researchers investigating a particular subject in information management and data insights can easily locate and search for topics and articles aligned with their research interests. Researchers Pathan & Prakash (2022) shared this same perception when applying LDA to analyze and understand hidden topics in unlabeled text documents. Their goal was to extract topics that represent aspects of their interest, based on the assumption that each document is a mixture of topics and that each topic is a statistical distribution of words.

It was observed that many documents in the studied dataset (a sample of scientific articles) exhibited mixed topic distributions (i.e., a tendency to integrate multiple conceptual domains), with few documents exclusively associated with a single topic. This is consistent with the multidisciplinary/interdisciplinary nature of contemporary scientific literature, in which works often integrate thematic areas with multiple approaches, as noted by Blei (2012), who reiterated that LDA models capture the latent structure of documents that naturally address multiple themes.

The authors Gibert et al. (2018) focused on expanding access to data science and DM for researchers and professionals in environmental sciences, promoting discussion of their use, and highlighting key DM concepts and techniques that could benefit environmental and natural resource science fields. In this process, DM was essential for detecting patterns in environmental databases, enabling the extraction of relevant information and helping identify the key parameters that manage and control the complex resource science field.

Researchers Dogan & Birant (2021) found that intelligent data analysis in the manufacturing industry is valuable because it generates new insights and provides a significant competitive advantage. They highlighted that essential features and structures are identified through data analysis; implicit knowledge, rules, and patterns present in the data are revealed through DM; furthermore, effective models are developed using ML to train computers to understand the behavior of a manufacturing system, learn from the data, and accurately predict future events.

A set of scientific articles was also explored by researchers Martins et al. (2023) as a valuable source of data for quantifying contaminants in food. These researchers also support the activities undertaken in this study by establishing relevant connections and identifying patterns in datasets, thereby providing a comprehensive view of the study object (food contamination). To achieve this, adaptations were made to the original FoodMines algorithm/protocol for literature mining (Hooton et al., 2020), which demonstrated strong performance metrics and good potential for application to a very large volume of articles.

Also, two studies on the applications of process mining are known, and they demonstrated the relevance of DM: one that improved the assessment and oversight of software development, covering

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

not just bug resolution but the entire software lifecycle (Kouzari et al., 2023), and another study that considered process mining functions as an intermediary between data science and operational management, providing organizations with a potent tool to enhance processes — such as concurrency, selection, and iteration — and achieve improved outcomes (Abrehdari, 2025).

Although numerous studies in the medical field examine DM techniques for predicting HIV infection risk, no single approach works across all types of datasets (Ge et al., 2024), which emphasizes that industry-specific DM analytical methods should continue to be overseen by trained human analysts.

The indiscriminate use of DM techniques, which was rightly criticized in the statistical literature in the 1990s with terms like "data dredging" to describe DM negatively without proper controls, can be a risky practice activity (Fayyad et al., 1996), often confused with genuine DM or exploratory analysis (Tukey, 1962, 1977)but distinguished by a lack of transparency that can easily lead to the discovery of meaningless and invalid patterns.

This kind of concern frequently emerges in literature discussions, emphasizing the importance of a unified set of guidelines to guarantee accurate analysis (Gibert et al., 2018) and avoid conflicting conclusions among data scientists, even when they analyze the same dataset.

By identifying opportunities to apply DM analytical techniques and ML models to the same dataset, researchers in Data, Information, and Management can examine how these analyses complement each other. This approach helps understand the field's evolution, pinpoint ongoing knowledge gaps, and emphasize that human interpretation, decision-making, and results should align with algorithmic automation.

5. Conclusion

This research applied analytical DM techniques — text mining and social network analysis — and ML methods — agglomerative hierarchical clustering, K-means clustering, and Latent Dirichlet Allocation (LDA) for topic modeling — to explore knowledge discovery in a collection of 287 scientific articles on information management and data insights. This work provided a more holistic view of the complementarity between these techniques and models: while DM focuses on discovering and interpreting patterns, ML focuses on building models that learn and generalize.

Among the main findings, it can be highlighted that: i) the *bigrams* (terms with two words) *artificial intelligence* and *machine learning* were the most frequently identified in the article titles, appearing 22 and 20 times, respectively; ii) five clusters were formed by the articles, and cluster C1 included the five most cited articles in this collection; iii) in a *co-occurrence* network of words identified in the article titles, *grouping 3* was formed by 23 nodes (28.6% of the network) – the largest number of nodes among all other groupings – and highlighted one of the central themes discussed in the journal, related to artificial intelligence, machine learning, and deep learning; iv) the *LDA topic modeling* generated six topics, with Topic 6, named *Machine Learning/Deep Learning and Predictive Modeling*, comprising 59 articles (20.6% of the sample) with a focus on technical and algorithmic aspects.

This article contributes to the Data, Information, and Management community by providing essential insights into the practical application of analytical approaches in DM and ML models, with a focus on extracting and discovering new knowledge.

The findings show that, although there were similarities and differences, combining techniques and models produced complementary understandings and revealed deeper insights into the studied data.

This work could motivate researchers in data, information, and management to deepen their understanding of the field, guide strategic academic development, and promote new knowledge and scientific advancement within the discipline.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

Author Contributions: Conceptualization, J.E.R.F. and E.R.F.; methodology, J.E.R.F. and E.R.F.; software, J.E.R.F.; validation, J.E.R.F. and E.R.F.; formal analysis, J.E.R.F.; investigation, J.E.R.F.; resources, J.E.R.F.; data curation, J.E.R.F.; writing—original draft preparation, J.E.R.F.; writing—review and editing, J.E.R.F.; visualization, J.E.R.F.; supervision, J.E.R.F.; project administration, J.E.R.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available from public sources.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI Artificial Intelligence
BoW Bag-of-Words
CPM Construction Project Management
DM Data Mining
EDM Educational Data Mining
IJIMDI International Journal of Information Management Data Insights
KDD Knowledge Discovery in Databases
LDA Latent Dirichlet Allocation
LSA Latent Semantic Analysis
ML Machine Learning
NLP Natural Language Processing
PFS Picture Fuzzy Set
SNA Social Network Analysis
STOs Security Token Offerings
TIM Technology and Innovation Management
TPB Theory of Planned Behavior
UTAUT Unified Theory of Acceptance and Use of Technology

References

1. Abbasi, A., Hossain, L., & Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, *6*(3), 403–412. https://doi.org/10.1016/j.joi.2012.01.002
2. Abrehdari, S. H. (2025). Leveraging the process mining technique to optimize data preparation time in a database used as an automated data delivery center. *MethodsX*, *15*, 103428. https://doi.org/10.1016/j.mex.2025.103428
3. Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, *1*(1). https://doi.org/10.1016/j.jjimei.2020.100004
4. Agrawal, R., Wankhede, V. A., Kumar, A., Luthra, S., & Huisingh, D. (2022). Big data analytics and sustainable tourism: A comprehensive review and network based analysis for potential future research. *International Journal of Information Management Data Insights*, *2*(2), 100122. https://doi.org/10.1016/j.jjimei.2022.100122
5. Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

6. Altuntas, F., Altuntas, S., & Dereli, T. (2022). Social network analysis of tourism data: A case study of quarantine decisions in COVID-19 pandemic. *International Journal of Information Management Data Insights*, *2*(2), 100108. https://doi.org/10.1016/j.jjimei.2022.100108
7. Altuntas, S., Selim, S., & Altuntas, F. (2022). A hierarchical clustering based panel data approach: A case study of regional incentives. *International Journal of Information Management Data Insights*, *2*(2). https://doi.org/10.1016/j.jjimei.2022.100098
8. Amorim, R. C. de, & Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, *324*, 126–145. https://doi.org/10.1016/j.ins.2015.06.039
9. Anugerah, A. R., Muttaqin, P. S., & Trinarningsih, W. (2022). Social network analysis in business and management research: A bibliometric analysis of the research trend and performance from 2001 to 2020. *Heliyon*, *8*(4), e09270. https://doi.org/10.1016/j.heliyon.2022.e09270
10. Bagirov, A. M., Aliguliyev, R. M., & Sultanova, N. (2023). Finding compact and well-separated clusters: clustering using silhouette coefficients. *Pattern Recognition*, *135*, 109144. https://doi.org/10.1016/j.patcog.2022.109144
11. Barbierato, E., Gatti, A., Incremona, A., Pozzi, A., & Toti, D. (2025). Breaking Away From AI: The Ontological and Ethical Evolution of Machine Learning. *IEEE Access*, *13*, 55627–55647. https://doi.org/10.1109/ACCESS.2025.3553032
12. Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks*. Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media, ICWSM 2009. https://doi.org/10.1609/icwsm.v3i1.13937
13. Biju, A. K. V. N., Thomas, A. S., & Thasneem, J. (2024). Examining the research taxonomy of artificial intelligence, deep learning & machine learning in the financial sphere - bibliometric analysis. *Quality and Quantity*, *58*(1), 849–878. https://doi.org/10.1007/s11135-023-01673-0
14. Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826
15. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993–1022. https://dl.acm.org/doi/10.5555/944919.944937-
16. Blondel, V. D., Guillaume, J.-L., & Lambiotte, R. (2024). Fast unfolding of communities in large networks: 15 years later. *Journal of Statistical Mechanics: Theory and Experiment*, *2024*(10), 10R001. https://doi.org/10.1088/1742-5468/ad6139
17. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
18. Bongini, P., Osborne, F., Pedrazzoli, A., & Rossolini, M. (2022). A topic modelling analysis of white papers in security token offerings: Which topic matters for funding? *Technological Forecasting and Social Change*, *184*. https://doi.org/10.1016/j.techfore.2022.122005
19. Cabana, E., Lillo, R. E., & Laniado, H. (2021). Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical Papers*, *62*(4), 1583–1609. https://doi.org/10.1007/s00362-019-01148-1
20. Coenen, F. (2011). Data mining: past, present and future. *The Knowledge Engineering Review*, *26*(1), 25–29. https://doi.org/10.1017/S0269888910000378
21. Cohen, J. (1988). Set Correlation and Contingency Tables. *Applied Psychological Measurement*, *12*(4), 425–434. https://doi.org/10.1177/014662168801200410
22. Crossref. (2024). *Crossref makes research objects easy to find, cite, link, assess, and reuse*. https://www.crossref.org
23. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
24. Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, *14*, 2349–2353. https://jmlr.org/papers/volume14/demsar13a/demsar13a.pdf

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

25. Dinh, T., Wong, H., Lisik, D., Koren, M., Tran, D., Yu, P. S., & Torres-Sospedra, J. (2025). Data clustering: a fundamental method in data science and management. *Data Science and Management*. https://doi.org/10.1016/j.dsm.2025.08.001
26. Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, 114060. https://doi.org/10.1016/j.eswa.2020.114060
27. Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). *Using Latent Semantic Analysis to improve access to textual information*. ACM Press. Proceedings of the SIGCHI - Conference on Human Factors in Computing Systems - CHI '88. https://doi.org/10.1145/57167.57214
28. Favaretto, J. E. R. (2015). Stage Level Measurement of Information and Communication Technology in Organizations. In *Academic Dissertation. FGV EAESP, Fundacao Getulio Vargas - Escola de Administracao de Empresas de Sao Paulo. Sao Paulo, Brazil.* (p. 105). https://doi.org/10.13140/RG.2.2.16844.82562
29. Favaretto, J. E. R. (2024a). *StopWords dataset: Integration of a set of stopwords in English and Portuguese - rev. 1. Zenodo*. Zenodo. https://doi.org/10.5281/zenodo.14176112
30. Favaretto, J. E. R. (2024b, October 21). *Artificial Intelligence (AI) explained by Venn diagram - rev. 1. Zenodo*. Zenodo. https://doi.org/10.5281/zenodo.17064700
31. Favaretto, J. E. R., & Meirelles, F. S. (2015). Nolan´S Stage Level Measurement of Information and Communication Technology in Modern Organizations. *46th Annual Southwest Decision Sciences Institute (SWDSI) Conference*, 410–418. https://www.favaretto.net/d/Favaretto_Meirelles__2015__nolans_stage_level_measurement_of_information_and_commnunication.pdf
32. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3). https://doi.org/10.1609/aimag.v17i3.1230
33. Francisco, E. D. R., Kugler, J. L., Kang, S. M., Silva, R., & Whigham, P. A. (2019). Beyond technology: Management challenges in the Big Data era. *Revista de Administração de Empresas*, *59*(6), 375–378. https://doi.org/10.1590/s0034-759020190603
34. Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph Drawing by Force-directed Placement. *Software - Practice and Experience*, *21*(11), 1129–1164. https://doi.org/10.1002/spe.4380211102
35. Gaviria-Marin, M., Merigó, J. M., & Baier-fuentes, H. (2019). Knowledge management: A global examination based on bibliometric analysis. *Technological Forecasting & Social Change*, *140*, 194–220. https://doi.org/10.1016/j.techfore.2018.07.006
36. Ge, Q., Lu, X., Jiang, R., Zhang, Y., & Zhuang, X. (2024). Data mining and machine learning in HIV infection risk research: An overview and recommendations. *Artificial Intelligence in Medicine*, *153*, 102887. https://doi.org/10.1016/j.artmed.2024.102887
37. Gelastopoulos, G., & Keramydas, C. (2025). A systematic review of text mining analytics for supply chain risk management using online data. *Supply Chain Analytics*, *12*, 100167. https://doi.org/10.1016/j.sca.2025.100167
38. Gephi. (2011). *Gephi Tutorial Layouts*. https://gephi.org/tutorials/gephi-tutorial-layouts.pdf
39. Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S. H., Rodríguez-Roda, I., & Holmes, G. (2018). Which method to use? An assessment of data mining methods in Environmental Data Science. *Environmental Modelling and Software*, *110*, 3–27. https://doi.org/10.1016/j.envsoft.2018.09.021
40. Godec, P., Pančur, M., Ilenič, N., Čopar, A., Stražar, M., Erjavec, A., Pretnar, A., Demšar, J., Starič, A., Toplak, M., Žagar, L., Hartman, J., Wang, H., Bellazzi, R., Petrovič, U., Garagna, S., Zuccotti, M., Park, D., Shaulsky, G., & Zupan, B. (2019). Democratized image analytics by visual programming through integration of deep models and small-scale machine learning. *Nature Communications*, *10*(1), 1–7. https://doi.org/10.1038/s41467-019-12397-x
41. Gupta, S. (2021). Deep learning based human activity recognition (HAR) using wearable sensor data. *International Journal of Information Management Data Insights*, *1*(2), 100046. https://doi.org/10.1016/j.jjimei.2021.100046
42. Hankar, M., Kasri, M., & Beni-Hssane, A. (2025). A comprehensive overview of topic modeling: Techniques, applications and challenges. *Neurocomputing*, *628*, 129638. https://doi.org/10.1016/j.neucom.2025.129638

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

43. Hooton, F., Menichetti, G., & Barabási, A.-L. (2020). Exploring food contents in scientific literature with FoodMine. *Scientific Reports*, *10*(1), 16191. https://doi.org/10.1038/s41598-020-73105-0
44. Hu, Y. (2006). Efficient, High-Quality Force-Directed Graph Drawing. *The Mathematica Journal*, *10*(1), 37–71.
45. Husna, A. ul, Ghasempoor, A., & Amin, S. H. (2024). A proposed framework for supplier selection and order allocation using machine learning clustering and optimization techniques. *Journal of Data, Information and Management*, *6*(3), 235–254. https://doi.org/10.1007/s42488-024-00127-y
46. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011
47. Kouzari, E., Sotiriadis, L., & Stamelos, I. (2023). Enterprise information management systems development two cases of mining for process conformance. *International Journal of Information Management Data Insights*, *3*(1), 100141. https://doi.org/10.1016/j.jjimei.2022.100141
48. Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights*, *1*(1), 100008. https://doi.org/10.1016/j.jjimei.2021.100008
49. Kushwaha, A. K., Kar, A. K., & Dwivedi, Y. K. (2021). Applications of big data in emerging management disciplines: A literature review using text mining. *International Journal of Information Management Data Insights*, *1*(2), 100017. https://doi.org/10.1016/j.jjimei.2021.100017
50. Lee, H., & Kang, P. (2018). Identifying core topics in technology and innovation management studies: a topic model approach. *Journal of Technology Transfer*, *43*(5), 1291–1317. https://doi.org/10.1007/s10961-017-9561-4
51. Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150–156. https://doi.org/10.1016/j.jesp.2017.09.011
52. Lian, Y., Hao, J., Zeng, W., & Luo, Q. (2025). A survey of visual insight mining: Connecting data and insights via visualization. *Visual Informatics*, *9*(4), 100271. https://doi.org/10.1016/j.visinf.2025.100271
53. Limongi, R., Brei, V. A., Almeida, V. de S. e, & Francisco, E. de R. (2025). 1+1 > 2: Integrating Analytical Techniques in the Age of AI. *Brazilian Administration Review*, *22*(4). https://doi.org/10.1590/1807-7692bar2025250230
54. Logan, A. P., LaCasse, P. M., & Lunday, B. J. (2023). Social network analysis of Twitter interactions: a directed multilayer network approach. *Social Network Analysis and Mining*, *13*(1), 65. https://doi.org/10.1007/s13278-023-01063-2
55. López-Meneses, E., López-Catalán, L., Pelícano-Piris, N., & Mellado-Moreno, P. C. (2025). Artificial Intelligence in Educational Data Mining and Human-in-the-Loop Machine Learning and Machine Teaching: Analysis of Scientific Knowledge. *Applied Sciences (Switzerland)*, *15*(2). https://doi.org/10.3390/app15020772
56. Lu, J. F., Tang, J. B., Tang, Z. M., & Yang, J. Y. (2008). Hierarchical initialization approach for K-Means clustering. *Pattern Recognition Letters*, *29*(6), 787–795. https://doi.org/10.1016/j.patrec.2007.12.009
57. Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139
58. Ma, T., Zhou, X., Liu, J., Lou, Z., Hua, Z., & Wang, R. (2021). Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies. *Technological Forecasting and Social Change*, *173*, 121159. https://doi.org/10.1016/j.techfore.2021.121159
59. Mahalanobis, P. C. (1930). On tests and measures of group divergence. *Journal and Proceedings, Asiatic Society of Bengal*, *XXVI*(4), 541–588.
60. Martins, Z. E., Ramos, H., Araújo, A. M., Silva, M., Ribeiro, M., Melo, A., Mansilha, C., Viegas, O., Faria, M. A., & Ferreira, I. M. P. L. V. O. (2023). From data to insight: Exploring contaminants in different food groups with literature mining and machine learning techniques. *Current Research in Food Science*, *7*, 100557. https://doi.org/10.1016/j.crfs.2023.100557

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

61. Mu, Y., Liu, X., & Wang, L. (2018). A Pearson's correlation coefficient based decision tree and its parallel implementation. *Information Sciences*, *435*, 40–58. https://doi.org/10.1016/j.ins.2017.12.059
62. Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, *31*, 274–295. https://doi.org/10.1007/s00357-014-9161-z
63. Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1). https://doi.org/10.1016/j.jjimei.2020.100007
64. Ordoñez-Avila, R., Salgado Reyes, N., Meza, J., & Ventura, S. (2023). Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review. *Heliyon*, *9*(3). https://doi.org/10.1016/j.heliyon.2023.e13939
65. Pathan, A. F., & Prakash, C. (2022). Cross-Domain Aspect Detection and Categorization using Machine Learning for Aspect-based Opinion Mining. *International Journal of Information Management Data Insights*, *2*(2), 100099. https://doi.org/10.1016/j.jjimei.2022.100099
66. Pereira, V., & Bamel, U. (2021). Extending the resource and knowledge based view: A critical analysis into its theoretical evolution and future research directions. *Journal of Business Research*, *132*, 557–570. https://doi.org/10.1016/j.jbusres.2021.04.021
67. Pereira, V. R., Penteado, P. P. B., & Favaretto, J. E. R. (2026). Analysis of the Financial Statements of the largest companies in the Brazilian agribusiness sector using Machine Learning techniques. *Revista Catarinense Da Ciência Contábil*, *25*, e3684. https://doi.org/10.16930/2237-7662202636842
68. Piatetsky-Shapiro, G. (2000). *Knowledge Discovery in Databases: 10 years after. SIGKDD Explorations. KDnuggets*. https://www.kdnuggets.com/gpspubs/sigkdd-explorations-kdd-10-years.html
69. Posit Team. (2026). *RStudio: Integrated Development Environment (IDE) for R. Posit Software, PBC, Boston, MA*. https://posit.co/products/open-source/rstudio
70. R Core Team. (2026). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org
71. Rejeb, A., Rejeb, K., Süle, E., Hassoun, A., & Keogh, J. G. (2025). Knowledge flows in industry 4.0 research: a longitudinal and dynamic analysis. *Journal of Data, Information and Management*, *7*(2), 123–145. https://doi.org/10.1007/s42488-025-00146-3
72. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
73. Sadeghi, B. (2022). Chatterjee Correlation Coefficient: A robust alternative for classic correlation methods in geochemical studies- (including "TripleCpy" Python package). *Ore Geology Reviews*, *146*, 104954. https://doi.org/10.1016/j.oregeorev.2022.104954
74. Saheb, T., Amini, B., & Kiaei Alamdari, F. (2021). Quantitative analysis of the development of digital marketing field: Bibliometric analysis and network mapping. *International Journal of Information Management Data Insights*, *1*(2), 100018. https://doi.org/10.1016/j.jjimei.2021.100018
75. Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia and Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864
76. ScienceDirect. (2026). *ScienceDirect - International Journal of Information Management Data Insights*. https://www.sciencedirect.com/journal/international-journal-of-information-management-data-insights
77. Scopus. (2024). *Scopus Source details - International Journal of Information Management Data Insights*. https://www.scopus.com/sourceid/21101081610
78. Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, *2*, 100003. https://doi.org/10.1016/j.nlp.2022.100003
79. Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796
80. Singh, S. S., Muhuri, S., Mishra, S., Srivastava, D., Shakya, H. K., & Kumar, N. (2024). Social Network Analysis: A Survey on Process, Tools, and Application. *ACM Computing Surveys*, *56*(8), 1–39. https://doi.org/10.1145/3648470

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 16 November 2026 https://doi.org/10.20944/preprints202611.1518.v1

81. Sonk, M., & Tunger, D. (2024). Trend mining with Orange – using topic modeling in futures research with the example of urban mobility. *European Journal of Futures Research*, *12*(1). https://doi.org/10.1186/s40309-024-00229-1
82. Spada, I., Giordano, V., Chiarello, F., Martini, A., & Fantoni, G. (2025). Text mining on green policies for integrating sustainability in higher education. *International Journal of Management Education*, *23*, 101126. https://doi.org/10.1016/j.ijme.2024.101126
83. Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 1–34. https://doi.org/10.1348/000711005X48266
84. Suha, S. A., & Sanam, T. F. (2023). Exploring dominant factors for ensuring the sustainability of utilizing artificial intelligence in healthcare decision making: An emerging country context. *International Journal of Information Management Data Insights*, *3*(1), 100170. https://doi.org/10.1016/j.jjimei.2023.100170
85. Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67. https://doi.org/10.1214/aoms/1177704711
86. Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.
87. Usai, A., Pironti, M., Mital, M., & Aouina Mejri, C. (2018). Knowledge discovery out of text data: a systematic review via text mining. *Journal of Knowledge Management*, *22*(7), 1471–1488. https://doi.org/10.1108/JKM-11-2017-0517
88. Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582
89. Verma, S., Sharma, R., Deb, S., & Maitra, D. (2021). Artificial intelligence in marketing: Systematic review and future research direction. *International Journal of Information Management Data Insights*, *1*(1), 100002. https://doi.org/10.1016/j.jjimei.2020.100002
90. Ward, J. M. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845
91. Ybañez, A., Ancheta, R., Evangelista, S. S., Aro, J. L., Maturan, F., Atibing, N. M., Selerio, E., Yamagishi, K., & Ocampo, L. (2023). How can we use machine learning for characterizing organizational identification - a study using clustering with picture fuzzy datasets. *International Journal of Information Management Data Insights*, *3*(1). https://doi.org/10.1016/j.jjimei.2023.100157
92. Yin, F., Lo, M. C., Mohamad, A. A., & Sin, K. Y. (2025). The impact of AI applications, information sharing, and supply chain resilience on agricultural supply chain performance. *Journal of Data, Information and Management*, *7*, 287–300. https://doi.org/10.1007/s42488-025-00155-2
93. Zheng, X., Le, Y., Chan, A. P. C., Hu, Y., & Li, Y. (2016). Review of the application of social network analysis (SNA) in construction project management research. *International Journal of Project Management*, *34*(7), 1214–1225. https://doi.org/10.1016/j.ijproman.2016.06.005

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.