algorithms

MDPI

Article

Mapping Web Scraping Research for Competitive Advantage: A Hybrid Bibliometric and Unsupervised Machine Learning Approach Using LDA Topic Modeling

Merve Vural Allaham

Department of Management Information Systems, Faculty of Economics, Administrative and Social Sciences, Istanbul Gelisim University, Istanbul 34310, Türkiye; mvural@gelisim.edu.tr

Abstract

This study examines the structural, semantic, and temporal evolution of strategy- and competition-oriented web scraping research, focusing on the legal and governance boundaries of data-driven competitive advantage. A purposively delimited corpus of 389 full-text articles indexed in the Web of Science and published between 2002 and 2025 was analyzed using bibliometric techniques and Latent Dirichlet Allocation (LDA), an unsupervised machine learning method for probabilistic topic modeling. Bibliometric analysis identified publication trends, keyword relationships, co-citation structures, and the visible conceptual organization of the field, while full-text LDA revealed ten latent themes. The findings show that data extraction, web crawling, social media content, and algorithmic structures remain central. However, the literature increasingly addresses market dynamics, consumer insights, organizational capabilities, digital skills, geospatial analysis, and data-driven decision-making. Comparing keyword clusters with full-text topics indicates thematic correspondence across several core areas, while demonstrating that full-text analysis captures contextual and application-oriented themes less visible in author keywords. Legal, ethical, and governance concerns constitute a distinct but relatively peripheral stream and show selective and uneven integration with strategically oriented themes, with stronger connections to data-extraction and social-media contexts than to market- and consumer-oriented themes. The study contributes methodologically by integrating full-text topic modeling with bibliometric network analysis and conceptually by identifying legitimate access, purpose of use, and organizational governance as boundary conditions of strategic value.

Keywords: web scraping; full-text topic modeling; Latent Dirichlet Allocation; strategic data infrastructure; dynamic capabilities; unfair competition; competitive intelligence

Academic Editor: Ariel Rosenfeld

Received: 4 September 2026
Revised: 30 September 2026
Accepted: 9 October 2026
Published: 10 October 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In the digital economy, competitive advantage increasingly depends on firms' capacity to collect data from their external environment, interpret such data, and translate it into strategic decision-making processes. Online data such as market, customer, competitor, platform, and regulatory data should not be viewed merely as fragmented digital content but rather as a raw material for competitive intelligence. The information systems literature increasingly emphasizes the centrality of data in business strategy, business

value creation, and business model innovation [1]. Similarly, unstructured data has been shown to offer significant opportunities for decision-making, environmental scanning, and the generation of managerial insight in business research [2]. Within this context, web scraping moves beyond its conventional definition as a technical method for extracting structured or unstructured data from websites through automated means. Instead, it becomes closely associated with firms' capacity to continuously monitor the external environment and support strategic insight. Recent studies indicate that web scraping, particularly when combined with AI-based models, enables deeper analysis of unstructured data and enhances its potential to create business value in decision-making processes [3]. Accordingly, the strategic value of web scraping does not stem solely from its capacity to collect data but from its ability to detect environmental signals at an early stage, process these signals analytically, and connect them to organizational decision-making mechanisms.

The central tension in the literature arises from the widespread use of web scraping alongside its limited theoretical positioning. Existing research primarily examines data extraction, crawling algorithms, social media analytics, AI-supported analysis, and legal or ethical risks [3,4]. Although recent studies demonstrate that scraped data can support organizational and sectoral analysis [5], the relationship between web scraping, strategic management, dynamic capabilities, and competitive intelligence remains insufficiently developed. Related information systems research suggests that digital data creates strategic value only when combined with analytical, human, and organizational capabilities [6–8]. Accordingly, this study examines web scraping as an external data infrastructure that may support competitive intelligence and environmental sensing [9]. A second tension concerns the legitimacy of the resulting advantage. Scraped data may support competitor monitoring, pricing, and market intelligence, but its acquisition and use may also raise concerns regarding ownership, unauthorised access, free riding, database rights, trade secrets, and platform rules. The unresolved issue is whether these concerns are treated as integral boundaries of strategic value creation or as a separate legal discussion.

The dominant treatment of web scraping in the literature as a technical method of data extraction limits its integration with strategic management and information systems theories, thereby giving rise to three major shortcomings. First, an integrated conceptual framework that connects technical, analytical, ethical, and strategic perspectives has yet to be fully developed. Second, the relationship between data obtained through web scraping and the notions of resources, capabilities, and strategic decision-making remains insufficiently theorized. Third, examining the field only at the level of keywords or abstracts may obscure the semantic transformations, thematic convergences, and intellectual shifts embedded in full texts, particularly the evolution of the literature from a narrow focus on technical data extraction toward strategic data infrastructure, competitive intelligence, environmental sensing, and governance. Although bibliometric analysis offers a powerful methodological approach for revealing the structural patterns of research fields, when used alone, it may remain limited in capturing the deeper semantic structure of texts and their latent thematic relationships [10]. Therefore, the web scraping literature should be examined not only through bibliometric indicators but also through full-text topic modeling.

A further gap concerns the limited integration between the strategic and legal streams of web scraping research. While one stream emphasizes the capacity of externally acquired web data to support competitive intelligence and organizational decision-making, another addresses privacy, intellectual property, platform terms, data ownership, and potentially unfair forms of competitive appropriation. It remains unclear whether and how these two streams are connected. In particular, the literature has not been

systematically mapped to determine how the pursuit of data-based competitive advantage is discussed alongside concerns relating to unfair competition and legal governance.

This study has a dual purpose. First, it maps the structural and semantic development of strategy- and competition-oriented web scraping research through a hybrid bibliometric and full-text LDA topic modeling approach. Second, it examines whether and how the literature addresses unfair competition and related legal-governance concerns arising from the use of web scraping to create information-based competitive advantage. The study therefore does not seek to represent the entire universe of web scraping research. Instead, it focuses on the intersection of web scraping, strategic value creation, competitive intelligence, and the legal boundaries of external data acquisition.

The central argument of this study is that web scraping may function as an external data acquisition mechanism supporting competitive intelligence and organizational sensing, but its strategic value cannot be evaluated independently of the legitimacy of data access, the purpose of use, data ownership, platform rules, and organizational governance. Web scraping is therefore conceptualized not as an inherently strategic or legitimate capability but as a contested data practice whose value depends on the analytical, organizational, and governance arrangements within which it is embedded. Accordingly, the research questions of this study are formulated as follows:

RQ1: What structural patterns are revealed by the development of the web scraping literature over time, its publication trends, and keyword relationships?

RQ2: What latent thematic structures and research clusters emerge in the web scraping literature at the full-text level?

RQ3: When the findings of bibliometric analysis and full-text topic modeling are evaluated together, how is web scraping positioned at the intersection of strategic value creation, competitive intelligence, and data governance?

RQ4: How are unfair competition and related legal-governance concerns represented within strategy-oriented web scraping research, and what boundaries do they establish for the use of web scraping as a source of competitive advantage?

RQ5: How can web scraping be conceptually positioned within the dynamic capabilities perspective, particularly in relation to the sensing dimension?

This study makes four contributions. First, it conceptualises web scraping as an external data acquisition infrastructure rather than merely a technical collection method. Second, it positions web scraping within the dynamic capabilities perspective, particularly as a mechanism that may support environmental sensing and competitive intelligence. Third, by integrating bibliometric analysis with full-text LDA topic modeling, it examines both the visible structural relationships and the latent semantic development of the field. Fourth, it incorporates unfair competition and legal governance into the strategic interpretation of web scraping by treating legitimate access, purpose of use, compliance, and organizational accountability as boundary conditions of data-based strategic value.

2. Literature Review

2.1. Definition and Application Areas of Web Scraping

Web scraping refers to the automated extraction of structured or unstructured website data and its transformation into an analysable format [11–13]. Its cost efficiency, timeliness, and customisability have increased its use in academic and commercial settings, including competitor-price monitoring, employee-review analysis, data enrichment, and skill forecasting [14–16]. However, dynamic webpages, access restrictions, API limits, and anti-scraping measures may affect its feasibility and reliability [17]. When used strategically, web scraping must also be considered in relation to data access, privacy, intellectual property, platform rules, and organizational responsibility [4,18]. The extraction of

economically valuable databases or licensed content may raise concerns regarding data ownership, free riding, unfair competition, and legal liability [19,20]. Its strategic value therefore depends not only on the usefulness of the data but also on how legitimately the data are accessed and used.

2.2. Research and Strategic Importance of Web Scraping

Web scraping enables organizations to monitor external environments through continuously updated information on prices, customers, competitors, platforms, and regulatory developments. Its relevance has increased as major platforms have restricted or monetised API access. Recent studies illustrate its use in retail research, competitor-price monitoring, market forecasting, regulatory analysis, and customer-review analytics [21–24]. Related computational research further demonstrates that web crawling can be combined with natural language processing to retrieve and transform unstructured online information into structured knowledge for downstream information-discovery and recommendation tasks [25]. When combined with appropriate analytical procedures, scraped data may therefore support competitive intelligence, market foresight, and customer insight. The central strategic issue is not access to data alone but whether externally acquired data are connected to organizational interpretation and decision-making.

2.3. Web Scraping as a Strategic Data Infrastructure

In the information systems and strategy literatures, data is increasingly regarded as a strategic resource. While the resource-based view conceptualizes data as one of the valuable resources through which organizations may achieve competitive advantage, recent information systems research emphasizes that data should be considered in close integration with business strategy and business model innovation [1,26]. However, the strategic value of data does not arise merely from its possession; rather, it depends on how data is collected, organized, analyzed, recontextualized, and embedded into decision-making processes [27]. From this perspective, web scraping can be understood as an external data acquisition mechanism that provides firms with a continuous flow of data from digital environments beyond organizational boundaries. Unlike one-time data fragments, scraped data may be reused and recontextualized across periods, purposes, and analytical settings [27,28].

In this study, strategic data infrastructure refers to the organizationally embedded mechanisms through which external digital data are continuously acquired, processed, interpreted, and connected to strategic decision-making. From a resource-based perspective, such infrastructure does not constitute a source of competitive advantage merely because data are available; rather, its value depends on the complementary analytical routines, technological resources, and organizational processes through which externally acquired data are transformed into actionable knowledge [29,30]. From a dynamic capabilities perspective, web scraping is particularly relevant to the sensing dimension because it can support the systematic monitoring of markets, competitors, customers, and other environmental signals [31,32]. Its contribution to seizing and transforming, however, depends on whether the acquired data are subsequently interpreted, integrated into managerial decisions, and translated into organizational action [30,31].

Although previous reviews and computational literature-mapping studies provide important foundations for the present research, their analytical scope leaves several issues unresolved. Barba et al. [3], which represents the closest review in substantive terms, examined 567 publications on the combined use of web scraping and AI-based models in business applications through bibliometric performance analysis and VOSviewer 1.6.20-based mapping of titles, abstracts, and keywords. Their study identified major application-oriented clusters and outlined future research directions, thereby demonstrating the

growing managerial relevance of web scraping and AI. However, it did not examine the latent thematic structure of the literature at the full-text level or explicitly investigate how web scraping is positioned in relation to competitive advantage, organizational capabilities, and the legal-governance boundaries of external data acquisition. In a related methodological stream, Zehir et al. [33] combined LDA topic modeling with keyword co-occurrence and temporal analysis to map the evolution of competitive strategy research. While this approach demonstrates the value of integrating latent semantic analysis with bibliometric networks, its unit of inquiry is the broader competitive-strategy literature, rather than web scraping as a specific mechanism of external data acquisition. Similarly, Semerikov and Mintii [34] integrated VOSviewer 1.6.20-based structural analysis with LDA topic modeling to examine both structural and semantic changes in a crisis-affected publication corpus, illustrating the methodological benefits of cross-method triangulation. Their analysis, however, was developed for a focused educational-psychology context and was not concerned with strategic data acquisition, competitive intelligence, or data-governance boundaries.

The present study extends these streams in both analytical scope and substantive focus. Rather than treating web scraping primarily as an application domain, an AI-related technique, or a component of a broader bibliometric landscape, it examines a purposively delimited corpus of 389 full-text publications at the intersection of web scraping, strategy, competition, and organizational decision-making. It combines bibliometric network analysis with full-text LDA topic modeling and subsequently examines document-level topic relationships to determine whether legal and ethical concerns are integrated with, or remain peripheral to, strategically oriented themes. The contribution therefore lies not simply in combining bibliometric analysis and topic modeling—an integration already demonstrated in prior studies—but in using these complementary methods to clarify how web scraping is positioned as an external data-acquisition mechanism associated with competitive intelligence and sensing, while simultaneously identifying legitimacy, purpose of use, and organizational governance as boundary conditions of its potential strategic value.

3. Theoretical Positioning

3.1. Resource-Based View: Web Scraping as a Mechanism for Accessing Strategic Data Resources

From a resource-based perspective, web scraping enables firms to access fragmented and often unstructured external data concerning markets, competitors, customers, platforms, and regulatory environments. However, the availability or accumulation of data does not itself create competitive advantage. External web data becomes strategically valuable only when it is combined with firm-specific analytical routines, interpretation, and decision processes that transform it into knowledge and competitive intelligence [27,29]. Web scraping should therefore be understood as an access mechanism rather than as a strategic resource in isolation.

3.2. Dynamic Capabilities: Web Scraping in Sensing, Seizing, and Transforming Processes

The dynamic capabilities perspective explains how organizations detect environmental change, evaluate opportunities, and adapt their resources and processes [31,32]. Within this framework, web scraping is most directly associated with sensing because it enables the systematic acquisition of external data concerning markets, competitors, customers, platforms, and regulatory developments. It may therefore operate as a micro-level mechanism supporting environmental monitoring and early signal detection. Connections with seizing and transforming are more indirect, as they depend on whether scraped data are analytically interpreted and subsequently embedded in decisions, innovation, or

organizational renewal [30,35]. Accordingly, the present study primarily positions web scraping as a sensing-related mechanism rather than as evidence of the complete dynamic capabilities process.

3.3. IT Capability and Digital Capability: Web Scraping as a Microfoundation of Digital Capability

Web scraping may also be positioned within a broader digital capability architecture. Data-driven capabilities require not only technological infrastructure but also the integration of data acquisition, processing, interpretation, and organizational use [36,37]. Web scraping does not constitute a digital capability in itself; rather, it represents a specific external data acquisition mechanism that may support market analytics, competitive intelligence, and data-driven decision-making.

3.4. Knowledge Economy, Unfair Competition, and Data Advantage Perspective

Web scraping may provide strategic advantage through access to competitors' prices, products, customer reviews, and platform activities. However, such advantage cannot automatically be regarded as legitimate. Unauthorised access, trade-secret violations, circumvention of platform rules, uncertain data ownership, and free riding complicate the legal and normative boundaries of data acquisition [4]. The strategic value of web scraping must therefore be evaluated together with the legitimacy of access, purpose of use, market effects, and organizational responsibility.

3.5. Knowledge and Competitive Intelligence Perspective

Scraped web data may expand organizations' knowledge of customers, competitors, markets, and sectoral change. When analyzed through methods such as sentiment and trend analysis, these data can support customer insight, market monitoring, and competitive intelligence [38,39]. Web scraping therefore functions as an input to intelligence generation rather than as an intelligence capability on its own.

4. Methodological Background

4.1. Topic Modeling and the LDA Approach

Latent Dirichlet Allocation (LDA) represents each document as a mixture of topics and each topic as a probability distribution over words, making it suitable for identifying latent thematic structures and organizing textual collections without predefined thematic categories [40–42].

4.2. Applications of Topic Modeling

Topic modeling has been applied to social media content, customer reviews, news, and other large-scale textual datasets to identify semantic patterns and research trends [39,43]. In this study, it is used to identify latent themes within the web scraping literature.

4.3. Rationale for Hybrid Bibliometric Analysis and Full-Text Topic Modeling

Bibliometric analysis reveals the visible structural configuration of a research field through publication trends, citation relationships, and keyword networks, but it may not fully capture thematic patterns embedded in full texts [10]. Full-text LDA topic modeling complements this structural perspective by identifying latent semantic dimensions. Their integration therefore enables the web scraping literature to be examined at both structural and semantic levels, while allowing keyword-based clusters to be compared with themes emerging from article content [17,41]. In particular, recent computational literature-mapping studies have demonstrated that integrating probabilistic topic modeling with

keyword co-occurrence and bibliometric network analysis can provide a more comprehensive representation of both latent thematic structures and visible conceptual relationships within a research field [33,34]. More broadly, recent computational bibliometric workflows increasingly complement conventional network-mapping tools with Python 3.12-based text processing and terminology standardization, illustrating the growing integration of algorithmic preprocessing with bibliometric research designs [44]. In this regard, while VOSviewer 1.6.20 captures keyword-based structural relationships, LDA extracts latent thematic dimensions from textual data, thereby enabling qualitative triangulation and deeper interpretation of findings [17]. Since this study adopts a multilayered analytical design, Table 1 presents the methodological framework by linking each analytical dimension with the corresponding method, tool, and research output.

From a science-of-science perspective, this hybrid framework enables the structural and semantic organization of a research domain to be examined simultaneously. By combining network-based bibliometric mapping with probabilistic full-text topic modeling, the study contributes to computational science-of-science research by demonstrating how visible scholarly relationships and latent thematic structures can be analyzed within an integrated framework.

Table 1. Multi-dimensional analytical framework: dimensions, methods, tools, and outputs.

Dimension	Method	Tool/Technique	Output
Data identification and screening	Systematic search and filtering	Web of Science, PRISMA-based screening	Final full-text corpus of 389 articles
Structural bibliometric mapping	Keyword co-occurrence analysis	VOSviewer 1.6.20	Thematic keyword clusters and intellectual structure of the field
Temporal development	Publication trend analysis	Python 3.12/descriptive statistics	Evolution of publication volume across years
Full-text semantic analysis	Latent Dirichlet Allocation topic modeling	Python 3.12/Gensim 4.4.0	10 latent topics with coherence score of approximately 0.52
Interpretive synthesis	Manual topic labeling and theoretical alignment	Researcher interpretation	Topic labels linked to web scraping, competitive intelligence, sensing, governance, and strategic data infrastructure
Strategic-theoretical interpretation	Cross-analysis of bibliometric and LDA findings	Manual alignment with RBV, dynamic capabilities, IT capability, and competitive advantage perspectives	Reframing web scraping as a strategic data infrastructure and sensing-related organizational mechanism
Topic integration analysis	Document-level topic co-occurrence analysis	Python 3.12; cosine similarity, joint salience, and lift	Integration pattern between the legal-ethical topic and the remaining LDA topics

5. Methodology

5.1. Data Collection

The identification and screening process was summarized using a PRISMA-based flow diagram. The articles included in the dataset were retrieved from the Web of Science Core Collection. The database search was conducted on 16 March 2026. No restrictions were applied to individual Web of Science indexes or document types at the search stage. The search strategy was designed to capture research at the intersection of web-based data extraction, strategy, competitive advantage, and data-driven organizational decision-making, while retaining the technical and governance-related dimensions of web scraping. Accordingly, the search was restricted to publications classified under the Web of

Science Research Areas of Business Economics, Computer Science, Information Science Library Science, Operations Research Management Science, Communication, Government Law, and Social Sciences Other Topics. These database-level restrictions were applied during the search stage, whereas language, full-text availability, text extractability, and substantive relevance were assessed subsequently during the screening process.

A total of two search queries were used. Query 1- ("web scraping" OR "data scraping" OR "screen scraping" OR "web crawling" OR "data harvesting") AND ("competitive advantage" OR competition OR strategy OR strategic) yielded 429 articles. Query 2- ("web scraping" OR "data scraping" OR "web crawling" OR "data harvesting") AND ("competitive advantage" OR "data advantage" OR strategy OR competition) yielded 394 articles. Since Query 1 and Query 2 produced highly similar results, there was a substantial overlap between the retrieved records, and duplicate articles were consolidated into single records. The search strategy was intentionally designed to capture the strategy- and competition-oriented segment of web scraping research rather than the entire technical literature on web crawling and information retrieval. Accordingly, technical subproblems such as page ranking, near-duplicate detection, and content classification were not included as independent search terms, as doing so would substantially broaden the corpus toward search-engine, indexing, and information-retrieval research beyond the analytical scope of the study. Such technical issues were not systematically excluded, however, and could enter the corpus when the corresponding studies also satisfied the web-scraping/crawling and strategy/competition search criteria.

After combining the results of these two queries, a total of 823 records were obtained. Following the removal of duplicate records, the full-text assessment process continued with 442 articles. To ensure consistency in natural language processing and semantic interpretation, only English-language documents were included in the analysis. Accordingly, eight articles written in Portuguese, Spanish, and Russian were excluded from the dataset. During the full-text analysis, 19 articles that the software identified as problematic and whose content could not be reliably read were also excluded. Upon inspection, these articles were found to consist predominantly of non-textual elements, such as figures and tables, and did not provide sufficient textual content for topic modeling. They were therefore excluded from the scope of the analysis. In addition, the full-text versions of 26 articles could not be accessed due to full-text availability limitations. As a result, a total of 389 articles were included in the final full-text analysis. Figure 1 summarizes the entire screening and selection process.

To retrieve the full texts of the articles, a Python 3.12-based query workflow was developed by combining DOI-based searches, the Crossref API for missing DOIs, and the Unpaywall API. In this process, the following libraries were used: requests for API communication and file downloading, pandas for reading and processing bibliographic data in Excel files, pathlib and os for file and directory management, re regular expressions for text cleaning and DOI normalization, and json for storing summary outputs. Using this automated procedure, only approximately one quarter of the full-text articles could be downloaded. The remaining full texts were therefore retrieved manually. All articles were obtained in March 2026.

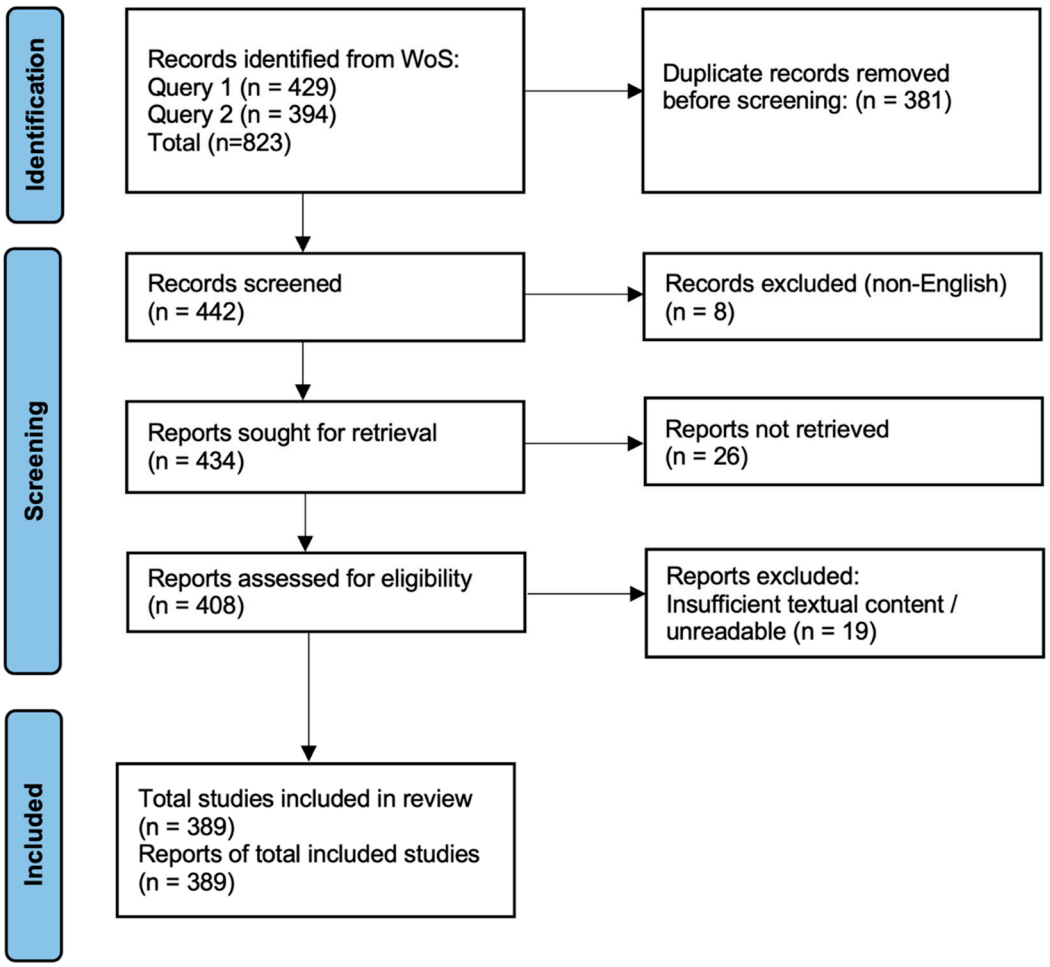


Figure 1. PRISMA-based data collection and screening process.

5.2. Data Preprocessing Process

PDF Extraction: Full-text data were extracted from PDF files using the pdfplumber library. Text extraction was conducted at the page level. Documents with no extractable text or with fewer than 100 characters were excluded from the dataset based on the criterion len(raw_text) > 100.

Data Cleaning: The textual data were preprocessed through a multi-stage cleaning procedure. To avoid bias arising from reference lists, reference sections identified through headings such as “references”, “bibliography”, and “cited studies” were detected and removed. In addition, URLs, email addresses, and DOI identifiers were eliminated. All textual expressions were converted to lowercase, and numerical values as well as non-alphabetic characters were removed. Excess whitespace was also normalized. These operations were performed using the re library for regular expressions.

Stopword and Noise Removal: The Natural Language Toolkit (NLTK) was used to remove standard stopwords in English and Spanish. In addition, domain-specific and academic noise terms were manually defined and excluded from the corpus. To further improve the robustness of the preprocessing procedure, the stopword detection mechanism of the spaCy library was also used during the token filtering stage.

Tokenization and Lemmatization: Tokenization and lemmatization were performed using the spaCy library. First, the texts were normalized and cleaned; they were then tokenized through spaCy, and each token was converted into its lemma form. To improve computational efficiency, the spaCy English language model (en_core_web_sm) was loaded with the parser and named entity recognition components disabled, as these

components were not required for tokenization and lemmatization. Several filtering criteria were applied to enhance data quality. Empty tokens, expressions included in the predefined stopword list, words shorter than three characters, and semantically weak verbs such as "be," "have," "do," and "use" were removed from the dataset. In addition, common functional words were filtered using spaCy's built-in stopword detection mechanism. Only tokens consisting exclusively of alphabetic characters were retained for analysis; expressions containing numbers, symbols, or mixed structures were excluded. Following preprocessing, documents were evaluated according to token count, and those containing fewer than 50 tokens were excluded from the analysis based on the criterion token_count ≥ 50.

N-gram Modeling (Bigrams and Trigrams): To capture meaningful multi-word expressions, bigram and trigram models were constructed using the Phrases module of the Gensim 4.4.0 library. Frequently co-occurring word pairs and triplets were identified according to predefined statistical thresholds (min_count = 5, threshold = 10). These expressions were then transformed into single tokens using the Phraser structure, such as competitive_advantage and data_analysis. The resulting n-grams were incorporated into the final token dataset used for topic modeling. The min_count and threshold values were selected to retain statistically meaningful multi-word expressions while reducing noisy or weak word combinations. The use of n-gram-based text processing is also increasingly incorporated into computational bibliometric workflows to reduce lexical inconsistency and improve the representation of meaningful multi-word structures in thematic analyses [45].

Corpus: Prior to estimating the LDA model, a dictionary and corpus structure were created using the Gensim 4.4.0 library. The dictionary represents the unique terms contained in the dataset, whereas the corpus consists of a bag-of-words representation of each document based on term frequencies. To improve model quality and reduce noise, extreme terms were filtered out. Specifically, terms appearing in fewer than three documents (no_below = 3) and terms occurring in more than 70% of documents (no_above = 0.7) were excluded from the analysis. This procedure was intended to minimize the adverse effects of both very rare and overly frequent terms on the model. The resulting corpus was then used as the input for the LDA model.

LDA Modeling: Latent Dirichlet Allocation (LDA) was selected because of its probabilistic structure and its interpretability in large-scale textual datasets. To determine the optimal number of topics, multiple LDA models were constructed using different topic configurations. The LDA models were trained using the Gensim 4.4.0 library. Specifically, models were estimated within a predefined range of topics, from 2 to 15, and the performance of each model was evaluated using the C_v coherence score. The C_v coherence measure was preferred because it combines indirect cosine similarity with normalized pointwise mutual information, making it particularly suitable for assessing semantic interpretability. To ensure consistency in coherence estimation across model runs, the number of processes was fixed at 1. The coherence score was used to evaluate model interpretability by measuring the semantic consistency of the high-probability words within each topic. The coherence values calculated for each model were then compared. The final number of topics was determined by considering both the C_v coherence scores and the substantive interpretability of the alternative solutions. In addition to the coherence values, the models were examined in terms of their top terms, representative documents, thematic overlap, topic size, and conceptual distinctiveness. The nine-topic solution combined conceptually distinct themes, whereas the twelve-topic solution fragmented several coherent themes into smaller and partially overlapping topics. The fifteen-topic solution resulted in even greater thematic fragmentation, producing several narrowly defined, weakly represented, and overlapping topics. By contrast, the ten-topic solution provided

a clearer, more parsimonious, and better-balanced thematic representation of the corpus and was therefore retained as the final model.

Using the selected model, topic distributions were obtained for each document, and the topic with the highest probability was assigned as the dominant topic for that document. To obtain complete topic distributions for each document, the minimum_probability = 0.0 setting was used during document-topic extraction. During model training, the parameters passes = 20, alpha = auto, and chunksize = 100 were applied. In addition, a fixed random_state = 42 was specified to ensure consistency and reproducibility of the results. This enabled a consistent comparison across different topic configurations. Overall, these parameter settings were selected to balance semantic coherence, model reproducibility, and computational efficiency across alternative topic models.

To assess the stability of the final ten-topic solution, the LDA model was re-estimated using five alternative random seeds (1, 7, 21, 100, and 2026), while all other model parameters and the corpus were held constant. Topic correspondence across runs was evaluated using Jensen–Shannon similarity, overlap among the top 20 terms, and correlations between document-level topic probabilities.

Visualization: To enhance the interpretability of the LDA results, visualizations were generated using the matplotlib and pyLDAvis 3.4.1 libraries. In particular, the intertopic distance map produced through pyLDAvis 3.4.1 was used to examine the spatial distribution of topics and their relative proximity to one another. This visualization enabled the assessment of topic overlap and the extent to which each topic was distinguishable from the others, that is, its level of topic distinctiveness. In addition, the keyword distribution of each topic was examined to interpret the semantic structure of the topics and support their meaningful labeling.

Topic Co-occurrence Analysis: To examine the degree to which the legal and ethical stream was integrated with or isolated from the broader thematic structure of the literature, an additional document-level topic co-occurrence analysis was conducted using the complete topic-probability distributions generated by the final LDA model. Because each document in LDA is represented as a probabilistic mixture of topics, integration was assessed using cosine similarity between the 389-dimensional document-level probability vector of Topic 8 (Legal and Ethical Issues in Data Scraping) and the corresponding probability vector of each of the remaining nine topics. Because minimum_probability = 0.0 was specified during document-topic extraction, a probability value was available for each of the ten topics in every document. The resulting 389 × 10 document-topic probability matrix was used as the input for the analysis. In addition, joint topic salience was examined using a probability threshold of 0.10, whereby two topics were considered jointly salient when both had probabilities of at least 0.10 within the same document. For each topic pair, the number of jointly salient documents and their proportion among documents with salient Topic 8 content were calculated. Co-occurrence lift was also calculated as the observed joint occurrence relative to the occurrence expected from the marginal prevalence of the two topics under statistical independence. Values above 1 indicate greater-than-expected co-occurrence, values around 1 indicate approximately proportional co-occurrence, and values below 1 indicate lower-than-expected co-occurrence. To assess the robustness of the findings to the selected probability threshold, the joint-salience analysis was repeated using thresholds of 0.05 and 0.15. All co-occurrence calculations were performed in Python 3.12 using the document-topic probability matrix exported from the final LDA model.

Cosine similarity: where θ_{ij} denotes the LDA probability of topic j in document i. The index i runs over the 389 documents.

$$\text{Cosine}(8,j) = \frac{\sum_{i=1}^{389} \theta_{i8}\,\theta_{ij}}{\sqrt{\sum_{i=1}^{389} \theta_{i8}{}^{2}}\sqrt{\sum_{i=1}^{389} \theta_{ij}{}^{2}}} \quad (1)$$

Co-occurrence lift: where τ is the topic-salience threshold (0.10 in the main analysis; 0.05 and 0.15 in the sensitivity analyses). A lift greater than 1 indicates greater-than-expected co-occurrence under statistical independence, a value near 1 indicates approximately proportional co-occurrence, and a value below 1 indicates lower-than-expected co-occurrence.

$$\text{Lift}(8,j) = \frac{P(T_8 \geq \tau, T_j \geq \tau)}{P(T_8 \geq \tau) \times P(T_j \geq \tau)} \quad (2)$$

6. Findings

6.1. Publications per Year

The articles reviewed were published between 2002 and 2025. In the figure below (Figure 2), the years of the studies and the number of publications made in that year are given to see when the studies gained momentum. The temporal distribution of publications reveals two distinct periods of development in the literature on web scraping and competitive advantage. Between 2002 and 2017, the number of publications was relatively low and stable. However, a significant upward trend has been observed since 2018. The peak in 2025 represents the highest level of scholarly interest, indicating that the subject has reached a stage of high visibility within the academic community.

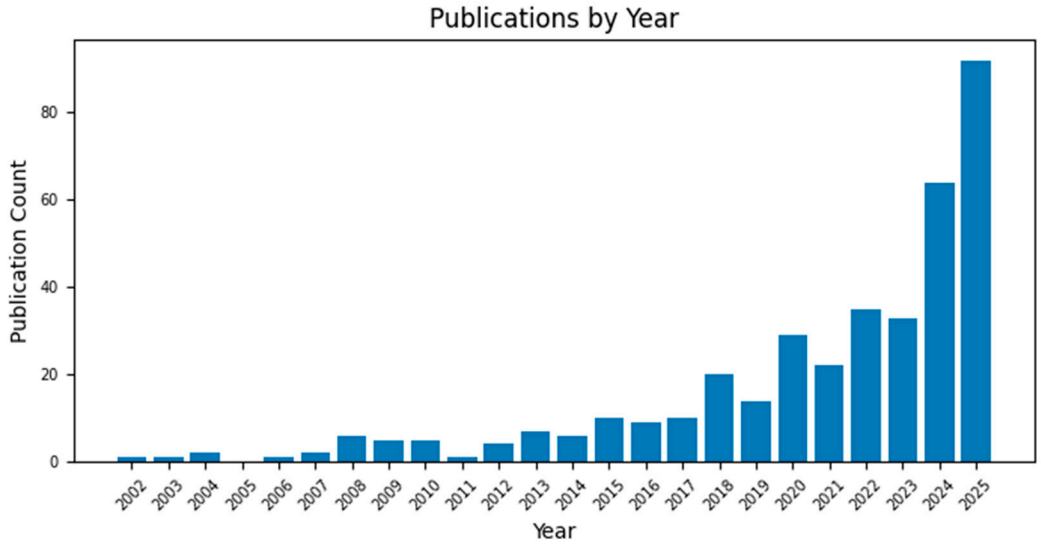


Figure 2. Annual Distribution of Publications.

6.2. Coherence Score Calculation

To determine the optimal number of topics for the LDA model, coherence scores were calculated across alternative topic configurations. Figure 3 presents the distribution of coherence values across different numbers of topics and illustrates their relative interpretability. Since coherence scores do not have a universally standardized threshold, the optimal model is generally selected based on the highest coherence value among alternative configurations, as this indicates stronger semantic interpretability [46]. Röder et al. (2015) [46] demonstrated that topic coherence measures are useful for evaluating semantic interpretability and comparing alternative topic solutions, but did not establish universal absolute cut-off values for distinguishing acceptable from unacceptable topic models. Accordingly, coherence is more appropriately interpreted comparatively across alternative models estimated from the same corpus and preprocessing procedure. Nevertheless,

practitioner-informed classifications have been proposed in subsequent applied research. For example, Yao et al. (2023) [47] classify C_v values between 0.50 and 0.55 as indicating medium coherence.

The results show that the coherence score reached its highest value at 10 topics (≈ 0.52). Thus, although the obtained value represents a moderate level of semantic coherence in absolute terms, it constitutes the highest relative coherence among the alternative topic configurations examined in this study. Lower numbers of topics, such as 2 to 7, yielded relatively lower coherence scores, indicating that these models may be overly generalized and unable to capture the thematic diversity embedded in the corpus. Conversely, excessively increasing the number of topics did not lead to continuous improvement in coherence, suggesting a trade-off between model complexity and interpretability.

Coherence scores varied across different topic configurations, with values of 0.4545 (2 topics), 0.4705 (3 topics), 0.4822 (4 topics), 0.4581 (5 topics), 0.4604 (6 topics), 0.4506 (7 topics), 0.4666 (8 topics), 0.4529 (9 topics), 0.5197 (10 topics), 0.5076 (11 topics), 0.5010 (12 topics), 0.4970 (13 topics), 0.5020 (14 topics), and 0.4940 (15 topics).

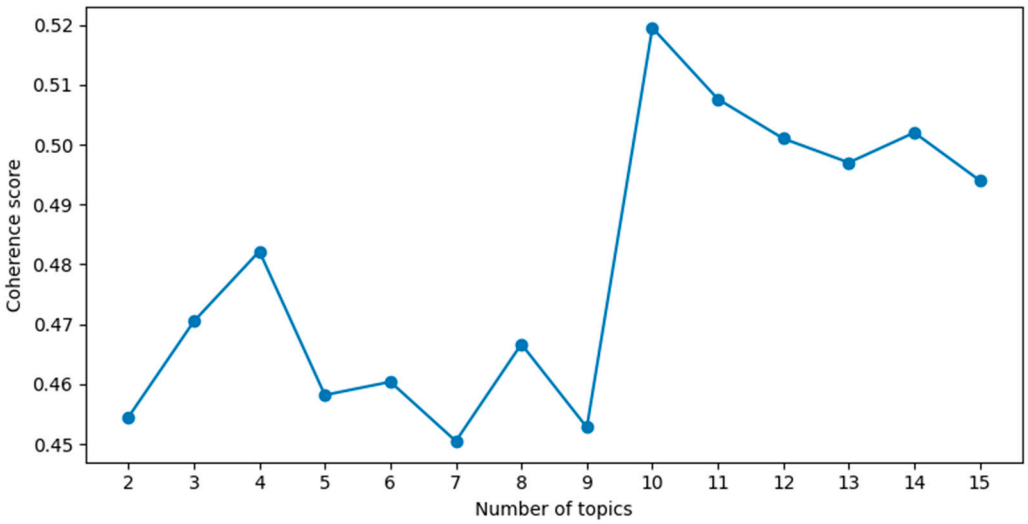


Figure 3. Topic Coherence Scores for Determining the Optimal Number of LDA Topics.

6.3. Topic Modeling

The selected ten-topic LDA solution identifies ten interpretable thematic structures within the corpus. The most dominant themes are related to "Data Extraction and Processing Techniques" (Topic 1), "Social Media Content and User Interaction" (Topic 2), and "Web Crawling and Algorithmic Structures" (Topic 4), suggesting a strong methodological and data-oriented orientation within the field. In addition, the presence of themes such as "Market Dynamics and Firm Performance" (Topic 3), "Organizational Capabilities and Digital Skills" (Topic 5), "Geospatial and Regional Data Analysis" (Topic 6), and "Sentiment Analysis and Consumer Insights" (Topic 7) indicates that web scraping is increasingly associated not only with technical data acquisition but also with organizational capabilities, data-driven decision-making, and multilayered contextual data. These findings suggest that web scraping plays a growing role in business analytics, managerial decision-making, and governance processes. At the same time, the theme "Legal and Ethical Issues in Data Scraping" (Topic 8) highlights the increasing relevance of regulatory and ethical dimensions of data use. Less dominant but conceptually important themes, such as "Urban and Tourism Analytics" (Topic 9) and "Sports Analytics Applications" (Topic 10), point to the diversification of application domains, including urban analytics, sports analytics, and geospatial data analysis. Topic labels were assigned through a structured

interpretive procedure that jointly considered the highest-probability terms, representative documents, the semantic context of these documents, topic distinctiveness, and inter-topic relationships. To enhance transparency and reduce arbitrary interpretation, Table 2 reports a representative document and an explicit interpretive rationale for each final topic label. Overall, the findings indicate that web scraping research has broadened beyond a predominantly technical focus and is increasingly discussed in relation to strategic applications, analytics, competitive intelligence, and digital transformation. The resulting topic labels are summarized in Table 2.

Table 2. LDA Topic Modeling Results: Dominant Topics, Key Terms, and Final Topic Labels.

Topic	Number of Articles Assigned to the Dominant Topic (n)	Tokens-Topic Terms	Final Topic Labels	Representative Document	Interpretive Rationale
Topic 1	83	(1, '0.009*"dataset" + 0.008*"website" + 0.006*"text" + 0.005*"web" + 0.005*"database" + 0.005*"data" + 0.005*"extract" + 0.004*"feature" + 0.004*"technique" + 0.004*"tool" + 0.004*"content" + 0.004*"classification"')	Data Extraction and Processing Techniques • data extraction • web scraping • data processing	Kim & Ha (2016), "Automated Discovery of Small Business Domain Knowledge Using Web Crawling and Data Mining" [48]	Represents how web crawling and data-mining techniques can transform unstructured online information into structured domain knowledge that supports small-business information needs and marketing-related decision-making.
Topic 2	73	(2, '0.008*"post" + 0.007*"content" + 0.005*"social_medium" + 0.005*"platform" + 0.004*"community" + 0.004*"user" + 0.003*"social" + 0.003*"individual" + 0.003*"support" + 0.003*"share" + 0.003*"relate" + 0.003*"health"')	Social Media Content and User Interaction • social media interaction • user-generated content • online communities	Li & Duan (2018), "Organizational Microblogging for Event Marketing: A New Approach to Creative Placemaking" [49]	Represents the use of social-media data to examine organizational communication, user engagement, and event-marketing activities, thereby representing the interaction-oriented use of online content.
Topic 3	39	(3, '0.015*"product" + 0.014*"price" + 0.011*"firm" + 0.010*"market" + 0.008*"company" + 0.006*"variable" + 0.005*"effect" + 0.005*"consumer" + 0.004*"level" + 0.004*"sustainability" + 0.004*"indicate" + 0.004*"change"')	Market Dynamics and Firm Performance • market analysis • pricing • firm performance	Castro et al. (2023), "Study of the Relationship Between Competitiveness and Digital Footprint Indicators in Valencian Wineries" [50]	Represents the market-oriented use of web-derived digital-footprint indicators by directly relating firms' online presence to competitiveness and economic-performance characteristics.

Topic 4	67	(4, ‘0.013*”page” + 0.012*”crawler” + 0.010*”crawl” + 0.009*”web_page” + 0.009*”algorithm” + 0.008*”set” + 0.007*”node” + 0.007*”link” + 0.007*”domain” + 0.006*”web” + 0.006*”content” + 0.005*”term”’)	Web Crawling and Algorithmic Structures • crawling algorithms • web structure • link analysis	Yu & Lin (2012), “Web Crawling and Filtering for On-line Auctions from a Social Network Perspective” [51]	Represents the combined use of web crawling, filtering, and network-based analysis to collect and organize online auction data for examining bidding behavior and electronic-market dynamics.
Topic 5	36	(5, ‘0.007*”university” + 0.007*”technology” + 0.007*”skill” + 0.006*”business” + 0.005*”management” + 0.005*”activity” + 0.005*”development” + 0.005*”phase” + 0.005*”service” + 0.005*”project” + 0.004*”company” + 0.004*”digital”’)	Organizational Capabilities and Digital Skills • organizational learning • digital capabilities • human capital	Pinski et al. (2023), “Executive AI Literacy: A Text-Mining Approach to Understand Existing and Demanded AI Skills of Leaders in Unicorn Firms” [52]	Represents the organizational-capability dimension by using digitally available information to identify AI-related knowledge and skills associated with leadership and emerging organizational competence requirements.
Topic 6	43	(6, ‘0.006*”image” + 0.005*”region” + 0.004*”respectively” + 0.004*”distribution” + 0.004*”year” + 0.004*”population” + 0.004*”area” + 0.004*”country” + 0.004*”dataset” + 0.003*”energy” + 0.003*”change” + 0.003*”observation”’)	Geospatial and Regional Data Analysis • demographic data • regional analysis • macro-level datasets	Laroche (2024), “Goodbye Monopoly: The Effect of Open Access Passenger Rail Competition on Price and Frequency in France on the High-Speed Paris–Lyon Line” [53]	Represents the use of geographically bounded market data to examine competitive entry, pricing, service frequency, and incumbent responses within a specific transport market.
Topic 7	27	(7, ‘0.019*”review” + 0.018*”network” + 0.017*”sentiment” + 0.015*”user” + 0.008*”brand” + 0.008*”word” + 0.007*”consumer” + 0.006*”comment” + 0.006*”customer” + 0.005*”medium” + 0.005*”content” + 0.005*”tweet”’)	Sentiment Analysis and Consumer Insights • sentiment analysis • customer reviews • brand perception	Chong et al. (2017), “Predicting Consumer Product Demands via Big Data: The Roles of Online Promotional Marketing and Online Reviews” [54]	Represents how online reviews and promotional information can be transformed into consumer insights and demand predictions, linking web-based data with market-oriented decision support.
Topic 8	8	(8, ‘0.027*”user” + 0.014*”app” + 0.013*”access” + 0.011*”scrap” + 0.010*”privacy” + 0.009*”court” + 0.009*”platform” +	Legal and Ethical Issues in Data Scraping • data privacy • legal regulations • platform policies	Parks (2022), “Unfair Collection: Reclaiming Control of Publicly Available Personal Information from Data Scrapers”	Represents the legal and ethical dimension of web scraping by addressing privacy, control, and legitimacy concerns

	0.009*"datum_scrap" + 0.009*"law" + 0.008*"public" + 0.008*"researcher" + 0.007*"linkedin"')		[55]	surrounding the collection and reuse of publicly accessible personal information.
Topic 9 8	(9, '0.021*"city" + 0.013*"tourism" + 0.012*"urban" + 0.010*"hotel" + 0.010*"area" + 0.010*"building" + 0.009*"airbnb" + 0.009*"tourist" + 0.007*"destination" + 0.007*"spatial" + 0.007*"housing" + 0.006*"beijing"')	Urban and Tourism Analytics • urban data • tourism economy • platform-based accommodation (Airbnb)	Jiang et al. (2025), "Configuring Placemaking Strategies Shaping Electronic Word of Mouth in Shanghai Shikumen Heritage Districts" [56]	Integrates web-derived social-media information with urban and tourism analysis to evaluate placemaking, visitor perceptions, electronic word of mouth, and destination-development strategies.
Topic 10 5	(10, '0.028*"team" + 0.024*"game" + 0.019*"sport" + 0.017*"match" + 0.016*"cricket" + 0.016*"play" + 0.015*"player" + 0.013*"score" + 0.011*"point" + 0.010*"ball" + 0.008*"world_cup" + 0.008*"competition"')	Sports Analytics Applications • sports analytics • performance data	Grassetti et al. (2021), "An Extended Regularized Adjusted Plus-Minus Analysis for Lineup Management in Basketball Using Play-by-Play Data" [57]	Represents a domain-specific sports-analytics application in which web-extracted performance data are converted into analytical information to support lineup evaluation and data-driven team-management decisions.

Note: n indicates the number of articles for which the corresponding topic had the highest document-level LDA probability (total N = 389). The asterisk (*) denotes multiplication between each token and its corresponding weight in the LDA topic representation and does not indicate statistical significance.

The topic solution also reveals a distinction between core strategic themes and more application-oriented themes. While several topics are directly associated with market analysis, organizational capabilities, consumer insights, and data-driven decision-making, others represent contextual domains in which web-based external data are collected and analysed. The application-oriented topics identified in the LDA solution can be theoretically interpreted through the sensing dimension of dynamic capabilities and the related concept of market sensing. Sensing refers to organizational processes through which changes, opportunities, threats, and other signals in the external environment are identified and interpreted [58]. Similarly, market-sensing capability concerns the systematic acquisition and interpretation of market information to improve organizational understanding of customers, competitors, and changing market conditions [59]. From this perspective, Topic 6 (Geospatial and Regional Data Analysis) represents a form of environmental sensing, as spatial, demographic, and regional data can provide contextual signals regarding locations, populations, demand structures, and changes in the external environment. Research on location intelligence similarly demonstrates that geospatial and regional information can support business location and strategic decision-making [60]. Topic 9 (Urban and Tourism Analytics) can be interpreted more specifically as sector-specific market sensing, because platform-, location-, demand-, and tourism-related digital data enable

organizations to detect market patterns, consumer movements, and changes in destination demand. Prior research has shown that geotagged and social-media-generated data can be transformed into decision-relevant information for strategic tourism planning and destination management [61]. Topic 10 represents a small and peripheral application cluster in the selected solution and should be interpreted cautiously because its emergence as a distinct topic was sensitive to random initialization. Sport-performance research has explicitly demonstrated parallels between competitive intelligence and the systematic collection and analysis of competitor and environmental information for strategic decision-making in sport [62]. Accordingly, these three topics are not interpreted as independent evidence of competitive advantage per se; rather, they represent contextual applications through which externally acquired web data can contribute to environmental sensing, market sensing, and competitive intelligence. Their retention therefore reflects the broader empirical manifestations of the sensing mechanism identified by the unsupervised topic model rather than an extension of the study's theoretical scope to unrelated application domains.

The repeated-run analysis showed that stability varied across topics. To provide a systematic assessment, Table 3 reports the mean Jensen–Shannon similarity, top-20 term overlap, and document-level topic-probability correlation obtained by matching each topic in the reference solution (random_state = 42) with its corresponding topic across five alternative random-seed estimations (1, 7, 21, 100, and 2026).

Table 3. Topic-Level Stability of the Final Ten-Topic LDA Solution Across Alternative Random Seeds.

Topic	Topic Labels	Mean JS Similarity	Mean Top-20 Term Overlap	Mean Document-Topic Correlation
Topic 1	Data Extraction and Processing Techniques	0.428	8.0	0.409
Topic 2	Social Media Content and User Interaction	0.543	10.8	0.650
Topic 3	Market Dynamics and Firm Performance	0.581	13.6	0.728
Topic 4	Web Crawling and Algorithmic Structures	0.643	16.6	0.856
Topic 5	Organizational Capabilities and Digital Skills	0.460	9.2	0.440
Topic 6	Geospatial and Regional Data Analysis	0.387	4.0	0.326
Topic 7	Sentiment Analysis and Consumer Insights	0.422	7.0	0.424
Topic 8	Legal and Ethical Issues in Data Scraping	0.424	12.0	0.597
Topic 9	Urban and Tourism Analytics	0.333	5.4	0.426
Topic 10	Sports Analytics Applications	0.232	0.0	−0.006

As shown in Table 3, Topic 4 exhibited the strongest stability across all three indicators, followed by Topics 3 and 2. Topic 8 also showed recurring semantic correspondence across alternative estimations, particularly in terms of top-term overlap and document-topic correlation. In contrast, the smaller application-oriented topics were less stable, with Topic 10 showing the greatest sensitivity to random initialization. These results indicate a comparatively stable thematic core alongside more variable peripheral application clusters and support cautious interpretation of the smaller topics.

The intertopic distance map shows that Topics 1, 2, 3, 4, 5, 6, and 7 are positioned relatively close to one another. This pattern suggests that, within the web scraping literature, data extraction, crawling, social media analytics, sentiment analysis, and various application domains are not sharply separated; rather, they form an intertwined knowledge structure. In contrast, Topics 8, 9, and 10 appear more distant from the remaining topics on the map. However, the overall configuration indicates that the literature is not characterized by a highly fragmented structure. Instead, it appears to cluster around a dense

central core of mutually reinforcing themes (see Figure 4). The full set of topic–term distributions is presented in Appendix A to enhance transparency and interpretability.

Notably, the presence of "classification" alongside "content," "feature," "accuracy," "score," and "algorithm" indicates that this topic also encompasses the classification and analytical processing of extracted web content, rather than representing data extraction alone.

The prominence of "page," "crawler," "crawl," "web_page," "algorithm," "node," "link," "network," "weight," and "engine" further shows that algorithmic crawling and page-level retrieval constitute an identifiable technical component of the corpus. However, specialised subproblems such as page ranking and duplicate-page detection did not emerge as sufficiently distinct latent themes in the final topic solution.



Figure 4. Intertopic Distance Map of the LDA Topic Model.

Note: The numbers (1–10) inside the circles represent the identifiers of the ten topics identified by the LDA model. Circle sizes indicate the relative prevalence of the topics in the corpus, while the distances between circles reflect their relative dissimilarities based on topic-word distributions.

6.4. Keyword Co-Occurrence

Keyword co-occurrence analysis provides a network-based representation of the conceptual structure of a literature by identifying terms that repeatedly occur together, while occurrence thresholds and total link strength can be used to retain the most structurally relevant concepts in the network [63]. In total, 1999 different keywords were used to classify studies in the sample. Figure 5 shows 132 terms referred to as keywords that are used with each other at least three times. The size of the nodes in the map shows how often each word is used. The length of the edge between two nodes indicates the strength of their relationship. The results indicate that "web scraping" is the central and most influential concept in the field, supported by strong connections with "big data," "social media," and "machine learning." This highlights the interdisciplinary nature of the literature, combining technical approaches with strategic and managerial perspectives (Table 4).

Table 4. Top Keywords by Co-Occurrence.

	Keyword	Occurrences	Total Link Strength
1	web scraping	59	138
2	big data	30	113
3	strategies	18	86
4	social media	29	83
5	impact	18	77
6	web crawling	39	66
7	sentiment analysis	17	56
8	management	11	52
9	machine learning	21	47
10	information	14	43

The clusters obtained from the keyword co-occurrence analysis are summarized in Table 5. Cluster 1 and Cluster 8 represent the technical and infrastructural dimensions of the literature, encompassing algorithms, crawling techniques, information retrieval, and data extraction processes. Cluster 2 and Cluster 7 are structured around concepts such as firm performance, strategy, innovation, and consumer behavior, thereby reflecting the business, marketing, and strategic management perspectives within the literature. Cluster 3 and Cluster 6 reveal application domains and data sources through themes such as social media, digital platforms, and user interaction, whereas Cluster 4 represents advanced analytical and methodological development through artificial intelligence, machine learning, and text analysis techniques. Cluster 5, through concepts such as ethics, access, and collaboration, points to the normative and governance-related dimension of the literature. The smaller and more specific clusters, namely Cluster 9 and Cluster 10, represent niche and context-specific research areas, such as data management infrastructures and education, indicating the diversity of application domains within the literature. Overall, this structure demonstrates that the web scraping literature is not merely a technical field; rather, it has evolved into an integrated research domain in which technical infrastructure, analytical methods, strategic use, and ethical dimensions are increasingly intertwined.

Interpreting co-occurrence communities as thematic clusters is consistent with network-based bibliometric applications in which densely connected keyword groups are used to reveal distinct but related conceptual domains within a research field [64].

Table 5. Thematic Clusters Derived from Keyword Co-Occurrence Analysis.

Cluster	Keywords	Theme
Cluster 1	algorithm, APIs (apis), behavior, classification, competition, data harvesting, design, e-commerce, energy, health, internet, market, model, network analysis, networks, patterns, price, prices, risk, search, UAV, web scraping	Algorithmic and Data Extraction Infrastructure
Cluster 2	china, discovery, entrepreneurship, firms, industry, innovation, labor, markets, models, nanotechnology, performance, quality, selection, strategies, strategy, teams, web mining, work	Strategic Management and Firm-Level Dynamics
Cluster 3	analytics, business intelligence, data analytics, data scraping, environment, facebook, hospitality, law, network, online, sentiment analysis, service, social media, social network analysis, social networks, tourism, twitter	Social Media Analytics and User-Generated Data
Cluster 4	artificial intelligence, artificial-intelligence, bibliometrics, deep learning, journalism, machine learning, natural language processing, news, nlp, python, technology, text analysis, text mining, topic modeling, web scraping, web scrapping, youtube	AI-Driven Text Analytics and Computational Methods

Cluster 5	access, challenges, collaboration, covid(COVID)-19, dimensions, ethics, focused crawler, implementation, information, knowledge, reputation, science, systems, trends	Ethical, Access, and Governance Challenges
Cluster 6	business, deep web crawling, digital platforms, engagement, extraction, fashion, influencers, instagram, latent dirichlet allocation, media, online social networks, spain, sustainability	Digital Platforms and Online Engagement Ecosystems
Cluster 7	big data, communication, consumer reviews, determinants, dynamics, impact, neural network, online reviews, perceptions, reviews, sales, word-of-mouth	Consumer Behavior and Market Intelligence Analytics
Cluster 8	firm performance, focused crawling, information retrieval, search engine, text, web, web crawler, web crawling	Information Retrieval and Web Crawling Systems
Cluster 9	content management system, data collection, data integration, data mining, optimization, wireless sensor networks	Data Management and Integration Technologies
Cluster 10	children, education, experiences, framework, management	Emerging and Context-Specific Application Domains

Figure 5 presents the network structure, showing the co-occurrence relationships among keywords. The concept of "web scraping", located at the center of the network, emerges as the most dominant theme in the literature and the keyword with the strongest relational strength. Its strong connections with concepts such as "big data," "machine learning," and "social media" indicate that the field is positioned at the intersection of technical data processing approaches, applied research domains, and analytical perspectives.

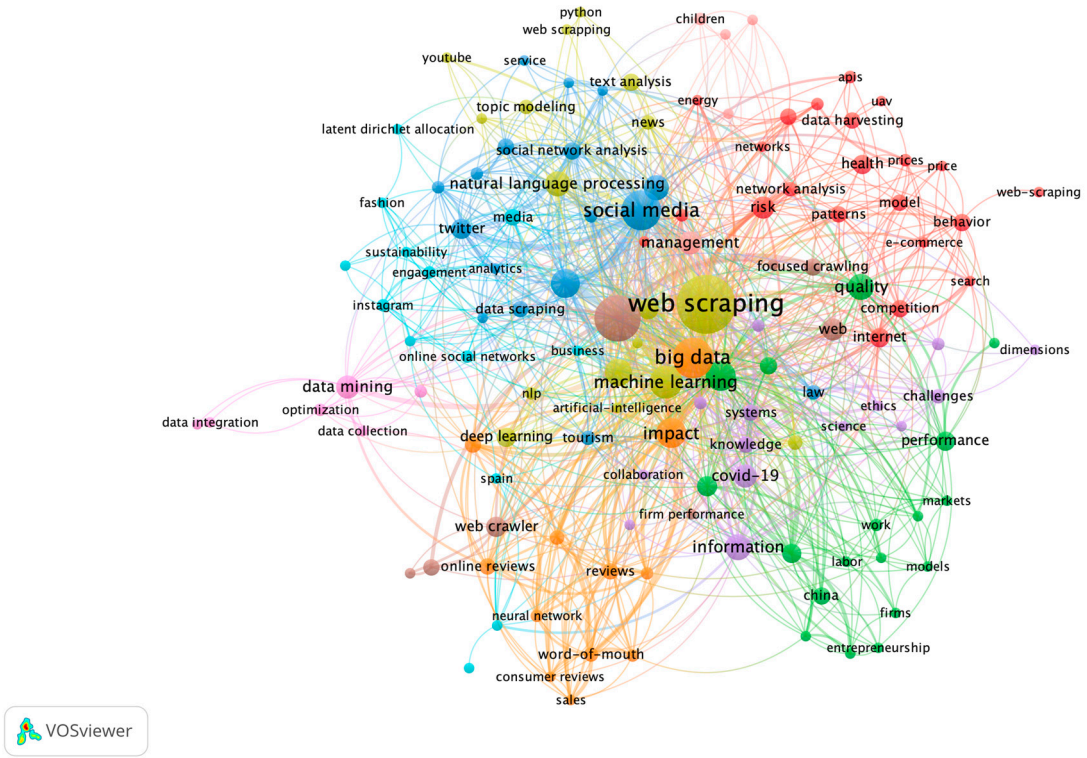


Figure 5. Keyword Co-Occurrence Network in Web Scraping Research.

Note: Node sizes represent keyword occurrence frequencies, while colors indicate clusters of related keywords. Links between nodes represent keyword co-occurrence relationships, with thicker links indicating stronger connections. The spatial proximity of nodes generally reflects their relatedness within the network.

6.5. Comparison of Keywords and Full-Text Topics

The comparison was treated as qualitative triangulation rather than formal validation. Correspondence was assessed by comparing the dominant terms, thematic scope, and substantive meaning of each VOSviewer 1.6.20 cluster with the highest-probability terms and interpretive content of the LDA topics. The comparison between the latent topics derived from the LDA model (Table 2) and the keyword co-occurrence clusters generated through VOSviewer 1.6.20 (Table 5) reveals both a substantial thematic correspondence and notable differences in analytical depth. Several core themes consistently identified by both methods point to well-established research streams within the web scraping literature. In particular, the technical foundation of the field centered on data extraction, web scraping, and algorithmic infrastructures is reflected in the alignment between Cluster 1 and Topic 1, as well as between Cluster 8 and Topic 4, both of which emphasize web crawling systems and algorithmic structures. Similarly, application-oriented themes such as social media analytics (Cluster 3 and Topic 2), consumer behavior and sentiment analysis (Cluster 7 and Topic 7), and ethical and legal issues (Cluster 5 and Topic 8) show a high degree of overlap, suggesting that these areas represent stable and relatively mature research streams. However, the correspondence was not one-to-one: Topics 5, 6, 9, and 10 had no clearly bounded counterparts in the keyword-cluster structure. However, the LDA results move beyond the scope of keyword-based clustering by revealing more context-specific and application-oriented themes, such as organizational capabilities, geospatial analysis, tourism analytics, and sports analytics, which are either less visible or more diffusely represented in the VOSviewer 1.6.20 clusters. This indicates that while VOSviewer 1.6.20 provides a structural and network-oriented overview of the research field based on keyword proximity, LDA captures deeper semantic patterns embedded in full-text data, thereby revealing latent thematic dimensions and contextual richness. Overall, the integration of these two methods highlights the multilayered nature of the web scraping literature, in which a strong technical core extends into diverse analytical applications and domain-specific use cases, while being increasingly shaped by ethical, legal, and governance-related concerns. In this respect, the bibliometric keyword co-occurrence network shows how the field is organized around visible conceptual structures, whereas LDA uncovers the deeper thematic layers underlying this surface structure. Interpreted together, these complementary approaches demonstrate that the literature is evolving simultaneously across technical, strategic, and regulatory dimensions.

6.6. Legal–Strategic Topic Co-Occurrence Analysis

To examine whether legal and ethical concerns constitute an isolated stream or are integrated with strategically oriented themes, document-level co-occurrence was analyzed using the complete LDA topic-probability distributions. Topic 8 (Legal and Ethical Issues in Data Scraping) exhibited a cosine similarity of 0.173 with Topic 5 (Organizational Capabilities and Digital Skills), 0.097 with Topic 7 (Sentiment Analysis and Consumer Insights), and 0.081 with Topic 3 (Market Dynamics and Firm Performance) (Table 6).

Table 6. Document-Level Co-occurrence Between the Legal-Ethical Topic and the Remaining LDA Topics.

Topic	Topic Name	Cosine Similarity with Topic 8 (Legal and Ethical Issues in Data Scraping)	Jointly Salient Documents (≥0.10)	% of Legal-Salient Documents	Lift
Topic 1	Data Extraction and Processing Techniques	0.279	29	65.9%	1.20

Topic 2	Social Media Content and User Interaction	0.246	22	50.0%	1.16
Topic 3	Market Dynamics and Firm Performance	0.081	7	15.9%	0.61
Topic 4	Web Crawling and Algorithmic Structures	0.116	8	18.2%	0.59
Topic 5	Organizational Capabilities and Digital Skills	0.173	11	25.0%	0.97
Topic 6	Geospatial and Regional Data Analysis	0.090	6	13.6%	0.41
Topic 7	Sentiment Analysis and Consumer Insights	0.097	6	13.6%	0.58
Topic 9	Urban and Tourism Analytics	0.036	1	2.3%	0.23
Topic 10	Sports Analytics Applications	0.010	0	0.0%	0.00

Note: Joint salience indicates that both Topic 8 and the corresponding strategic topic had document-level probabilities ≥ 0.10. Lift values above 1 indicate greater-than-expected co-occurrence, values around 1 indicate approximately independent occurrence, and values below 1 indicate lower-than-expected co-occurrence.

Using a topic-probability threshold of 0.10, 44 documents contained salient legal and ethical content. Among these documents, 11 (25.0%) also exhibited salient Topic 5 content, 6 (13.6%) exhibited salient Topic 7 content, and 7 (15.9%) exhibited salient Topic 3 content. Overall, 22 of the 44 documents (50.0%) contained at least one of the three strategically oriented topics at the same threshold. However, the pairwise co-occurrence lift values were 0.97, 0.58, and 0.61, respectively, indicating that joint occurrence was not greater than would be expected from the marginal prevalence of these topics in the corpus.

The broader topic-overlap pattern further contextualizes this result. Topic 8 showed greater similarity with Topic 1 (Data Extraction and Processing Techniques; cosine = 0.279) and Topic 2 (Social Media Content and User Interaction; cosine = 0.246) than with any of the three strategically oriented topics. Thus, legal and ethical concerns are not completely isolated from the broader web scraping literature, but their integration is uneven: they are more closely embedded in discussions concerning data extraction and platform-based or user-generated content than in themes explicitly concerned with market performance, consumer insight, or organizational capability.

Sensitivity analyses produced a consistent pattern. At a probability threshold of 0.05, 40 of the 60 documents (66.7%) with salient legal content also contained at least one strategically oriented topic, whereas at the more restrictive threshold of 0.15, this number decreased to 8 of 31 documents (25.8%). The corresponding overall lift decreased from 0.982 at the 0.05 threshold to 0.520 at the 0.15 threshold. Taken together, these results indicate selective and uneven rather than complete integration between the legal-governance and strategic streams.

6.7. Temporal Analysis

To account for the substantial increase in publication volume over time, temporal keyword trends were evaluated using relative rather than raw frequencies. The corpus was divided into three periods: 2002–2012, 2013–2019, and 2020–2025. For each keyword kk in period pp, relative frequency was calculated as $RF_{k,p} = (n_{k,p}/N_p) \times 100$, where $n_{k,p}$ is the number of publications containing the keyword and N_p is the total number of publications in that period. Temporal change was expressed as the percentage-point difference between the last and first periods.

The publication-volume-normalized analysis reveals a differentiated temporal pattern. The relative prevalence of keywords such as web scraping, social media, machine learning, natural language processing, artificial intelligence, and sentiment analysis increased between the first and last periods, whereas technically oriented terms such as web crawling and focused crawling declined substantially in relative prevalence. These findings are consistent with a gradual broadening of the literature from crawling- and infrastructure-oriented research toward a wider combination of data-intensive, AI-supported, and application-oriented themes (Figure 6).

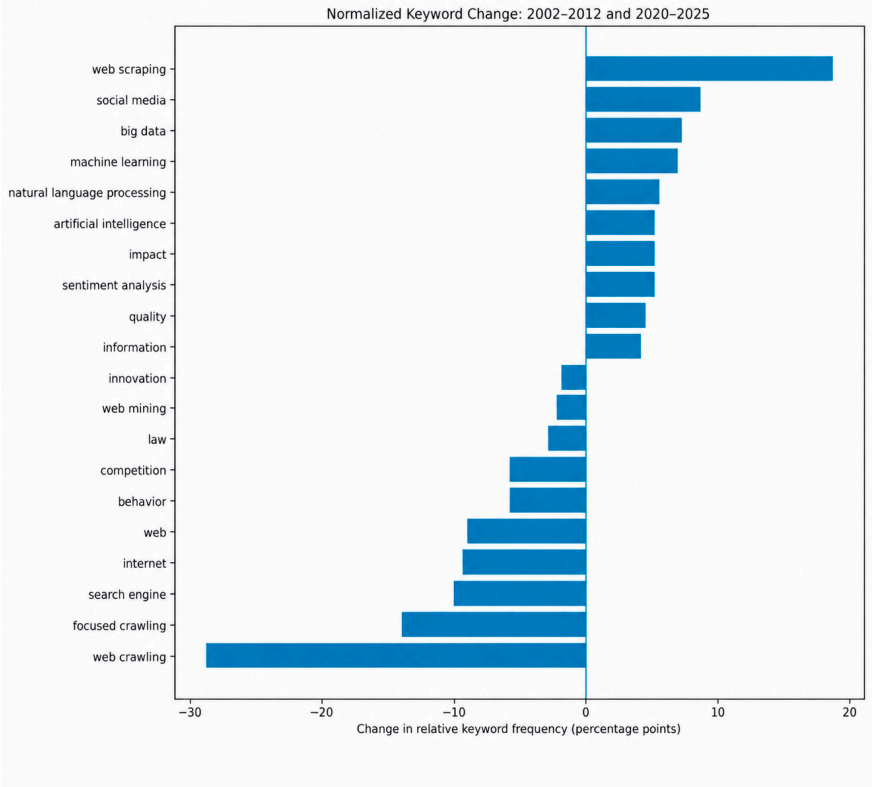


Figure 6. Changes in Publication-Volume-Normalized Relative Keyword Frequencies Between 2002–2012 and 2020–2025.

The publication-volume-normalized temporal trends indicate that several AI- and analytics-related keywords became more visible in the later years of the corpus, particularly after 2018. Relative frequencies of terms such as artificial intelligence, sentiment analysis, and social media increased during the 2020–2025 period, although year-to-year fluctuations remain evident. Considered together with the period-based comparison presented in Figure 6, these patterns are consistent with a gradual broadening of web scraping research toward data-intensive, AI-supported, and application-oriented themes. Figure 7 therefore reflects changes in the relative prominence of keywords rather than changes in absolute publication counts.

Overall, these findings indicate a gradual thematic broadening within the sampled strategy- and competition-oriented corpus, from a predominantly technical focus toward a more application-driven and intelligence-oriented research landscape. This pattern is consistent with the increasing visibility of big data, artificial intelligence, and platform-based analytical approaches within the literature.

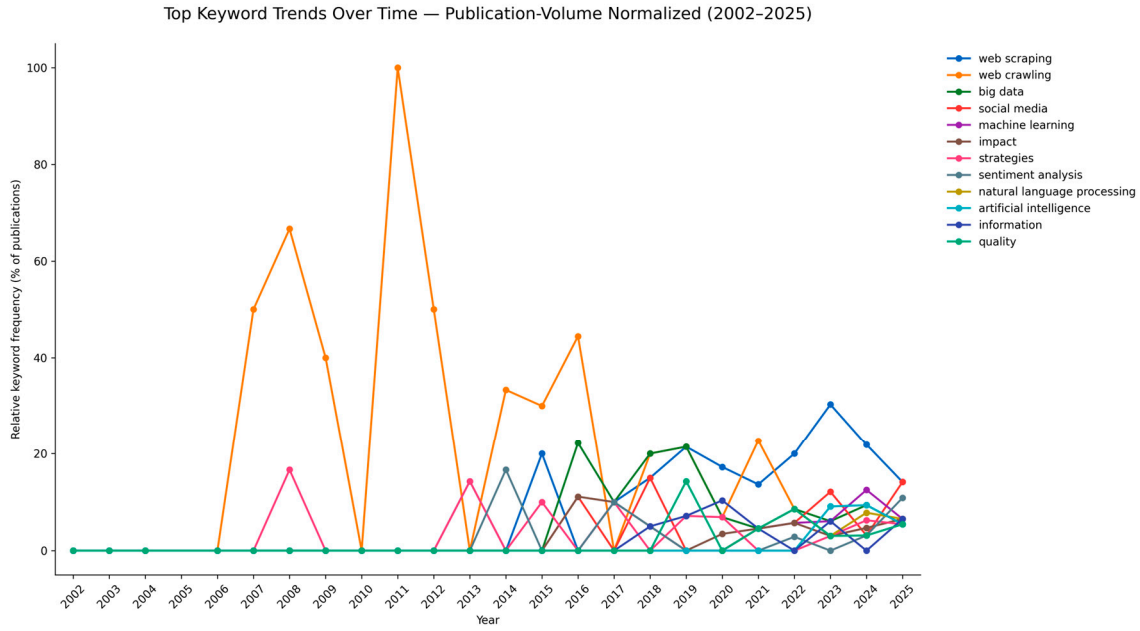


Figure 7. Publication-Volume-Normalized Relative Keyword Trends Over Time.

6.8. Co-Citation—Cited References

Co-citation analysis examines how frequently pairs of references are cited together and is commonly used to reveal the intellectual structure of a research field. Co-citation analysis is used to analyze and understand the intellectual structure of a discipline [65]. When the same reference pairs are cited together in publications on a subject, research clusters are formed [66]. Co-citation analysis is a suitable method for mapping the research area as a whole. As a result of the analysis, the resources that are linked are cited at least five times together. Of the 17,127 cited references, 33 meet the threshold (Table 7).

Table 7. Most Cited Publications and Their Total Link Strengths.

Rank	Publication	Research Focus/Contribution	Citations	Total Link Strength
1	Brin & Page (1998) [67]	Introduced the architecture of a large-scale web search engine and established PageRank as a foundational link-based ranking approach.	11	35
2	Heydon & Najork (1999) [68]	Presented Mercator, a scalable and extensible web crawler architecture designed for large-scale web data collection.	12	30
3	Chakrabarti et al. (1999) [69]	Developed a focused crawling approach for discovering topic-specific web resources more efficiently than general crawling.	11	27
4	Olston & Najork (2010) [70]	Provided a comprehensive review of web crawling foundations, technical challenges, and emerging trends in information retrieval.	8	25
5	Cho et al. (1998) [71]	Proposed URL ordering strategies to improve crawling efficiency and prioritize the retrieval of valuable web pages.	7	23
6	Adar et al. (2009) [72]	Examined the dynamic nature of web content and how changes in online information affect crawling and data updating strategies.	7	21

7	Cho & Garcia-Molina (2003) [73]	Developed methods for estimating how frequently web pages change, supporting more effective crawler scheduling and refresh policies.	7	21
8	Diligenti et al. (2000) [74]	Introduced focused crawling using context graphs to improve the discovery of relevant web pages within specific topical domains.	6	19
9	Pinkerton (1994) [75]	Presented early experiences with the WebCrawler system, contributing to the development of web search and crawling technologies.	5	18
10	Boldi et al. (2004) [76]	Proposed UbiCrawler, a scalable and fully distributed web crawler architecture for large-scale web exploration.	6	17

6.9. Co-Citation—Cited Sources

Cited sources are interesting for evaluating the co-citations of journals, the selection of journals and the collection of journals that serve publication-oriented researchers on a particular subject. Journal co-citation networks reveal the macro structure of the scientific discipline through the analysis of journal titles. The evaluation of the co-citation analysis of web scraping and competition publications in terms of the published source is presented in Table 8. In total, 9220 cited sources were analyzed. The cited sources with the highest citation counts and strongest link strengths include PLOS ONE (121 citations; TLS = 1797), Journal of Business Research (83 citations; TLS = 1711), Sustainability (104 citations; TLS = 1251), and Scientific Reports (37 citations; TLS = 1190). These are followed by domain-relevant outlets such as Journal of Retailing and Consumer Services, Journal of Cleaner Production, and Management Science, as well as interdisciplinary and high-impact sources like PNAS and the preprint repository arXiv.

The prominence of journals such as PLOS ONE, Scientific Reports, and arXiv indicates a strong orientation toward data-driven, computational, and interdisciplinary research practices within the field. At the same time, the presence of journals such as Journal of Business Research, Management Science, and Journal of Business Ethics highlights the growing importance of strategic, managerial, and ethical perspectives. Notably, the significant representation of Sustainability and Journal of Cleaner Production suggests that web scraping research is increasingly linked to sustainability-oriented studies, including environmental monitoring, sustainable consumption, and resource optimization. This indicates a shift toward leveraging web-based data for addressing broader societal challenges and supporting data-driven sustainability transitions. Furthermore, application-oriented outlets such as Journal of Retailing and Consumer Services and sustainability-focused journals suggest that web scraping research is increasingly applied in domains such as consumer behavior, digital markets, and sustainability analytics.

Table 8. Top Cited Sources by Citations and Total Link Strength.

	Name of Journal	**Citations**	**Total Link Strength**
1	Plos One	121	1797
2	Journal of Business Research	83	1711
3	Sustainability	104	1251
4	Scientific Reports	37	1190
5	Journal of Retailing and Consumer Services	71	1019
6	Journal of Cleaner Production	68	1002
7	arXiv	149	995
8	Management Science	48	978

9	Journal of Business Ethics	39	860
10	Proceedings of the National Academy of Sciences of the USA (PNAS)	42	845

7. Discussion

The most important finding of this study is that the web scraping literature can no longer be viewed as a narrowly defined field organised solely around technical data extraction and crawling architectures. The findings indicate that the field has gradually expanded towards a broader data infrastructure logic that obtains data from the external environment, interprets this data through analytical processes, and links it to decision-making mechanisms. The increase in publications over the years, the strong intersection of the keyword network with concepts such as "web scraping," "web crawling," "big data," "social media," "machine learning," and "strategies," and the emergence of the dimensions of "data extraction and data processing techniques," "social media content and user interaction," "market dynamics and firm performance," and "organizational capabilities and digital skills" in the full-text analysis suggest that this expansion does not merely reflect a superficial change in terminology. More specifically, the field is not losing its technical origins. Rather, its technical core is increasingly being reinterpreted within strategic and managerial contexts of use. This study argues that web scraping can be conceptualized not only as a data collection technique but also as an external data acquisition infrastructure that supports the systematic monitoring of the digital environment beyond firm boundaries [5].

The acceleration of scholarly interest after 2018 may be associated with the wider adoption of big data analytics, advances in AI-supported methods, the growing centrality of platform economies, and changes in formal data access channels [77]. This growth may therefore reflect not only methodological interest but also an increasing need to access external data within platform-controlled ecosystems [3].

When the topic modeling results and keyword clusters are considered together, it becomes evident that the field has not entirely abandoned its technical core but has added layers of application, analysis, and governance to this foundation. The prominence of topics such as "Data Extraction and Processing Techniques" and "Web Crawling and Algorithmic Structures" indicates that the field remains connected to its computational infrastructure. In contrast, the emergence of topics such as "Market Dynamics and Firm Performance," "Sentiment Analysis and Consumer Insights," and "Organizational Capabilities and Digital Skills" suggests that a more institutional and managerial discourse concerning the strategic value of data has developed within the literature.

The combined findings connect the field's computational foundations with its emerging managerial and strategic applications. The strategic value of web scraping does not arise from data extraction alone but from the processing, interpretation, and integration of scraped data into organizational decision-making. This interpretation is consistent with process-based resource-based explanations, according to which data derive value from their organizational use rather than from access or ownership alone [1,27,30,31]. A technically reliable data acquisition infrastructure may therefore be considered a prerequisite for strategic use.

The third and theoretically strongest interpretation emerges from the connection established with the dynamic capabilities perspective. The findings show a particularly strong conceptual alignment between web scraping and the sensing dimension of dynamic capabilities. The identified themes involve practices such as monitoring environmental change, market signals, prices, user reviews, and platform dynamics. However, the study does not empirically demonstrate that web scraping improves firm performance or directly creates competitive advantage. Rather, web scraping may be understood as an

external data acquisition mechanism that supports sensing when combined with appropriate analytical and managerial capabilities.

Another finding of the study is that, although ethical and legal dimensions are not among the quantitatively most dominant themes, they have become visible as a relatively distinct thematic area. This finding provides a qualified answer to whether the literature is concerned with the possibility that web scraping may generate competitive advantage through legally or normatively problematic means. The literature does address this issue, but it does not yet constitute a dominant or fully integrated research stream. Legal discussions are more frequently organised around privacy, access rights, platform rules, intellectual property, data ownership, and organizational responsibility than around a direct theorisation of unfair competition. Unfair competition therefore appears less as an established central framework and more as a boundary problem surrounding data-based competitive advantage.

Although ethical and legal issues are not among the most dominant themes, they constitute a distinct area of the literature. Discussions focus primarily on privacy, access rights, platform rules, intellectual property, data ownership, and organizational responsibility rather than on the direct theorisation of unfair competition. Unfair competition therefore appears mainly as a boundary problem surrounding data-based competitive advantage. The legitimacy of such an advantage depends on the nature of the data, the method of access, its subsequent use, and compliance with legal, contractual, and organizational boundaries [4].

Previous studies have highlighted the use of web scraping for data access, price monitoring, customer review analysis, social media analysis, and sectoral observation [38,39]. The present findings confirm these technical and analytical applications but extend them by positioning web scraping within a strategic data infrastructure perspective. By linking external data acquisition with organizational capabilities, digital skills, and competitive intelligence, the study integrates themes that remained fragmented in earlier technically oriented reviews [6,8,9]. The diversity of application areas further indicates that web scraping has evolved into a general-purpose external data acquisition mechanism rather than remaining confined to a single research domain.

The intertopic distance map shows that Topics 1–7 are positioned close to one another and that data extraction, crawling, social media, sentiment analysis, and application areas are not sharply differentiated. By contrast, the topics of law and ethics, tourism and urban analytics, and sports analytics are positioned further apart. The study indicates that the overall structure is not highly fragmented, but is organised around a dense central cluster. This result suggests strong thematic transitions among technical, analytical, and applied research streams within the web scraping literature. Technical tools are generally not examined in isolation; rather, they are considered together with specific data sources and analytical purposes.

The findings provide complementary evidence for the five research questions. For RQ1 and RQ2, the bibliometric and full-text topic modeling results show that the sampled strategy- and competition-oriented web scraping literature retains a technical core centred on data extraction, crawling, and algorithmic structures, while expanding towards artificial intelligence, social media analysis, market dynamics, organizational capabilities, consumer insights, and governance. This reflects thematic broadening rather than the replacement of technical research by strategic research. For RQ3, the findings position web scraping at the intersection of strategic value creation, competitive intelligence, and data governance. Although data extraction remains its operational foundation, the strategic value of scraped data depends on their integration with analytical interpretation, organizational processes, market intelligence, and governance arrangements. Web scraping may therefore contribute to a broader strategic data infrastructure, but its strategic value is

conditional on how the data are interpreted, governed, and incorporated into decision-making. For RQ4, the document-level co-occurrence analysis qualifies the characterization of legal-governance concerns as a peripheral stream. These concerns are not completely isolated from strategically oriented research; rather, their integration is selective and uneven. Topic 8 (Legal and Ethical Issues in Data Scraping) exhibited substantially greater overlap with Data Extraction and Processing Techniques (cosine = 0.279) and Social Media Content and User Interaction (0.246) than with Organizational Capabilities and Digital Skills (0.173), Sentiment Analysis and Consumer Insights (0.097), or Market Dynamics and Firm Performance (0.081). Moreover, none of the three strategic pairings produced a lift above 1 at the 0.10 salience threshold. This pattern suggests that legal and ethical concerns are more frequently embedded in discussions concerning the acquisition, access, and platform context of web data than in research explicitly focused on its strategic outcomes. Consequently, legitimacy should not be interpreted as an empirically established moderator of strategic value in the present study; rather, the findings support its conceptual positioning as a boundary condition whose integration into the strategic web-scraping literature remains incomplete. For RQ5, the findings do not empirically demonstrate that web scraping enhances dynamic capabilities or firm performance. However, themes such as environmental monitoring, competitor and customer tracking, price movements, platform activities, and regulatory change indicate strong theoretical alignment with the sensing dimension of dynamic capabilities. Web scraping may therefore support organizational sensing, while its translation into seizing, transforming, and improved performance depends on complementary analytical, managerial, and organizational capabilities that require future empirical investigation.

Finally, the study's hybrid methodological approach revealed that the themes identified through full-text analysis did not fully correspond to the author keywords. While keyword co-occurrence analysis presented the visible, networked, and structural aspects of the field, full-text LDA identified contextual and application-oriented themes that were less visible in the keywords. Niche application areas such as organizational capabilities, tourism, and sports, as well as geospatial and regional data uses, became more visible at the full-text level. This indicates that attention should be paid not only to the concepts through which the field is defined but also to the application and theoretical contexts addressed in the full texts of the studies. Author keywords are limited in number and selected strategically. Therefore, a study's data context, application area, or secondary theoretical orientation may not be represented in its keywords. Full-text analysis can capture this otherwise invisible semantic layer.

7.1. Theoretical Implications

The primary theoretical contribution of this study is to conceptualise web scraping not merely as a technical method but as a component of external data acquisition infrastructure. This perspective extends resource-based interpretations focused on data access or ownership by showing that strategic value depends on analytical processing, reuse, and integration into organizational decision-making. Second, the study positions web scraping as a potential micro-foundation of organizational sensing. While sensing refers to the identification of environmental opportunities and threats, the external data acquisition mechanisms supporting this process are not always specified. Web scraping addresses this gap by enabling the systematic collection of data from the external digital environment. However, the findings do not directly establish links with seizing or transforming. These links depend on the analytical and organizational processes through which externally acquired data are interpreted and used, providing a research agenda for examining the complementary capabilities required to extend sensing into subsequent dynamic capability processes. Third, the findings suggest that the IT capability and digital

capability literatures should recognize external data acquisition as a distinct capability layer. Existing research often emphasises analytical infrastructure, data processing, and digital transformation, while giving less explicit attention to the systematic acquisition of external digital data. Web scraping should not be treated as a digital capability in itself but as a potential micro-foundation supporting external data access and environmental intelligence. Its contribution depends on its integration with technical infrastructure, organizational processes, and governance arrangements. Finally, the study adds a normative dimension to the concept of strategic data infrastructure. The strategic value of web scraping cannot be assessed solely through data volume or analytical capacity but must also account for the legitimacy of access, purposes of use, platform rules, and organizational responsibility. Governance should therefore be treated as a constitutive element of strategic data infrastructure rather than as a secondary constraint.

7.2. Practical Managerial Implications

From a managerial perspective, organizations should treat web scraping not as a fragmented data collection activity but as an institutionalised process linked to specific decision problems. Managers should first determine which external data sources support which business objectives—for example, using price data for competitive analysis and customer reviews for product or service improvement. Without such alignment, scraping may generate data and costs without improving decisions. Organizations should also establish a governance and compliance framework before data collection begins. Public accessibility does not necessarily eliminate legal, contractual, or institutional risks. A basic assessment should cover platform access rules, API alternatives, privacy, licensing, intellectual property, and documentation requirements. Although this requires initial investment, it can reduce legal, reputational, and access-related risks. In the medium term, scraping should be integrated with analytical systems and decision routines, as data create value only when they are cleaned, combined, interpreted, and used. This requires collaboration among data engineering, analytics, and business teams. For example, pricing data can be converted into indicators of competitor behaviour, while customer reviews can be classified into actionable product or service categories. Such integration requires technical infrastructure and qualified personnel. In the long term, web scraping should be institutionalised as part of the organization's digital capability architecture. This involves not only software development but also processes for data quality, source documentation, system maintenance, human oversight, and legal review. Sustainable implementation depends less on extracting more data than on ensuring that the data are reliable, explainable, and usable for decision-making. The findings do not support a universal performance benefit from web scraping. Implementation should therefore be gradual and limited to contexts in which external digital data are decision-relevant and adequate analytical and governance capabilities exist. Otherwise, poor data quality, access restrictions, maintenance costs, and misinterpretation may outweigh the potential benefits.

7.3. Limitations and Future Research

The first limitation is that the dataset was restricted to Web of Science and English-language publications. Studies indexed in other databases or published in different languages may reveal alternative thematic patterns, particularly across different legal and platform contexts. Future research could use multiple databases and multilingual corpora to assess the robustness of the findings. The second limitation concerns full-text access and extraction. Some publications were excluded because their full texts were unavailable or could not be processed reliably, potentially underrepresenting studies with visual or non-standard formats. Future research could reduce this bias through broader institutional access, alternative acquisition methods, and multimodal document-processing

techniques. Third, LDA-based topic modeling involves researcher interpretation, particularly in topic selection and labelling. The coherence score of approximately 0.52 indicates reasonable interpretability, but the model should be viewed as one plausible thematic representation of the corpus. Although the core thematic structure was comparatively stable across repeated runs, several smaller application-oriented topics were more sensitive to random initialization. Accordingly, the absence of a distinct topic should not be interpreted as evidence that specific technical issues are absent from web scraping research generally. In particular, although content classification is represented within Topic 1 and several page-level and algorithmic crawling concepts appear within Topic 4, specialised issues such as page ranking and duplicate-page detection did not emerge as sufficiently prominent or semantically concentrated to constitute independent latent topics within the purposively delimited strategy- and competition-oriented corpus. Future studies could benchmark LDA against alternative approaches such as non-negative matrix factorization, neural topic models, or contextualized topic models, as comparative evidence suggests that alternative topic-modeling architectures may differ in their ability to generate coherent, diverse, and interpretable thematic structures [78]. Time-sliced analyses and multiple coders could further strengthen the robustness of topic interpretation. Finally, the literature-mapping design does not test whether and under what conditions web scraping supports dynamic capabilities or firm performance.

The purposive delimitation of the corpus should also be considered when interpreting the observed integration patterns. Because the search strategy was intentionally designed to capture the strategy- and competition-oriented segment of web scraping research, the resulting corpus does not represent the entire web scraping literature. Legal, ethical, or highly technical studies that do not explicitly intersect with the selected strategic and competition-related terms may therefore be underrepresented. Consequently, the observed legal–strategic co-occurrence patterns should be interpreted as characterizing relationships within this intentionally delimited corpus rather than as evidence of the integration structure of the broader web scraping literature.

Future empirical research could examine the proposed conceptual relationships through complementary quantitative, qualitative, and longitudinal designs. A survey-based study could operationalize web-scraping maturity, analytical capability, governance capability, sensing capability, and firm performance as distinct organizational constructs. Web-scraping maturity could capture the extent to which organizations systematically collect, update, integrate, and reuse externally acquired web data, whereas analytical capability could reflect the organization's ability to transform such data into interpretable insights. Governance capability could encompass formal policies, access controls, compliance mechanisms, accountability structures, and procedures governing the legitimate use of externally acquired data. These constructs could then be linked to sensing capability, defined as the organization's ability to identify and interpret relevant market, competitor, customer, platform, and regulatory signals. Structural equation modeling or related multivariate approaches could be used to test whether web-scraping maturity is directly associated with sensing capability and whether this relationship is mediated by analytical and governance capabilities. Subsequent models could examine whether sensing capability, in turn, is associated with decision quality, responsiveness, innovation-related outcomes, or firm performance.

Complementary multiple-case studies could provide a process-level understanding of how scraped data are actually transformed into organizational sensing, seizing, and transforming routines. Such studies could compare organizations that differ in industry, digital maturity, data-governance practices, and intensity of web-scraping use in order to identify the organizational conditions under which externally acquired web data become strategically useful. Particular attention could be given to how firms select data sources,

validate scraped information, combine it with internal data, interpret emerging signals, and translate those signals into pricing, marketing, innovation, or competitive-response decisions. Case-based evidence could also help identify circumstances in which web scraping fails to generate value because of poor data quality, weak analytical capability, organizational resistance, legal constraints, or inadequate governance.

Longitudinal research would be particularly valuable for examining temporal ordering and reducing the limitations of cross-sectional evidence. Repeated observations could assess whether increases in web-scraping maturity precede improvements in sensing capability, decision quality, or performance outcomes over time, rather than merely being correlated with them at a single point. Longitudinal or quasi-experimental designs could also examine organizational changes before and after the adoption of systematic web-scraping practices, the introduction of new governance mechanisms, or changes in platform-access conditions. Together, these approaches would allow future research to move beyond the conceptual relationships identified in the present literature-mapping study and directly test when, how, and under what organizational conditions web scraping contributes to dynamic capabilities and firm-level outcomes.

8. Conclusions

This study examines the web scraping literature through hybrid full-text topic modeling and bibliometric analysis, revealing that, while the field retains its focus on technical data extraction, it is increasingly being associated with a broader strategic data infrastructure perspective. The findings indicate that technical dimensions such as data extraction and crawling continue to constitute the core of the field; however, they are now increasingly linked to market dynamics, consumer insights, organizational capabilities, social media analytics, and ethical and legal governance. This conceptual framework positions web scraping not merely as a technical tool but as a potential mechanism for transferring external digital data into organizational information processes.

From a literature perspective, the study establishes a clearer theoretical bridge between web scraping and the strategic management, information systems, and competitive intelligence literatures. From a conceptual and practical perspective, the findings suggest that the strategic value of externally acquired data may depend not solely on access and collection but also on governance, analytical interpretation, and integration into organizational decision-making processes. Therefore, the strategic importance of web scraping depends less on the volume of data collected than on the extent to which such data can be transformed into a responsible, interpretable, and decision-useful information infrastructure. More broadly, the study demonstrates the value of integrating bibliometric networks with full-text machine learning as a computational approach to science-of-science research.

Funding: This study was supported by the Scientific Research Projects Coordination Unit of Istanbul Gelisim University [project number: SKB-2025-200].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The search strategies, bibliographic list of included studies, derived analytical outputs, and analysis code supporting the findings of this study are available from the corresponding author upon reasonable request. The full texts of the analyzed articles are not redistributed because they remain subject to the copyright and licensing conditions of their respective publishers.

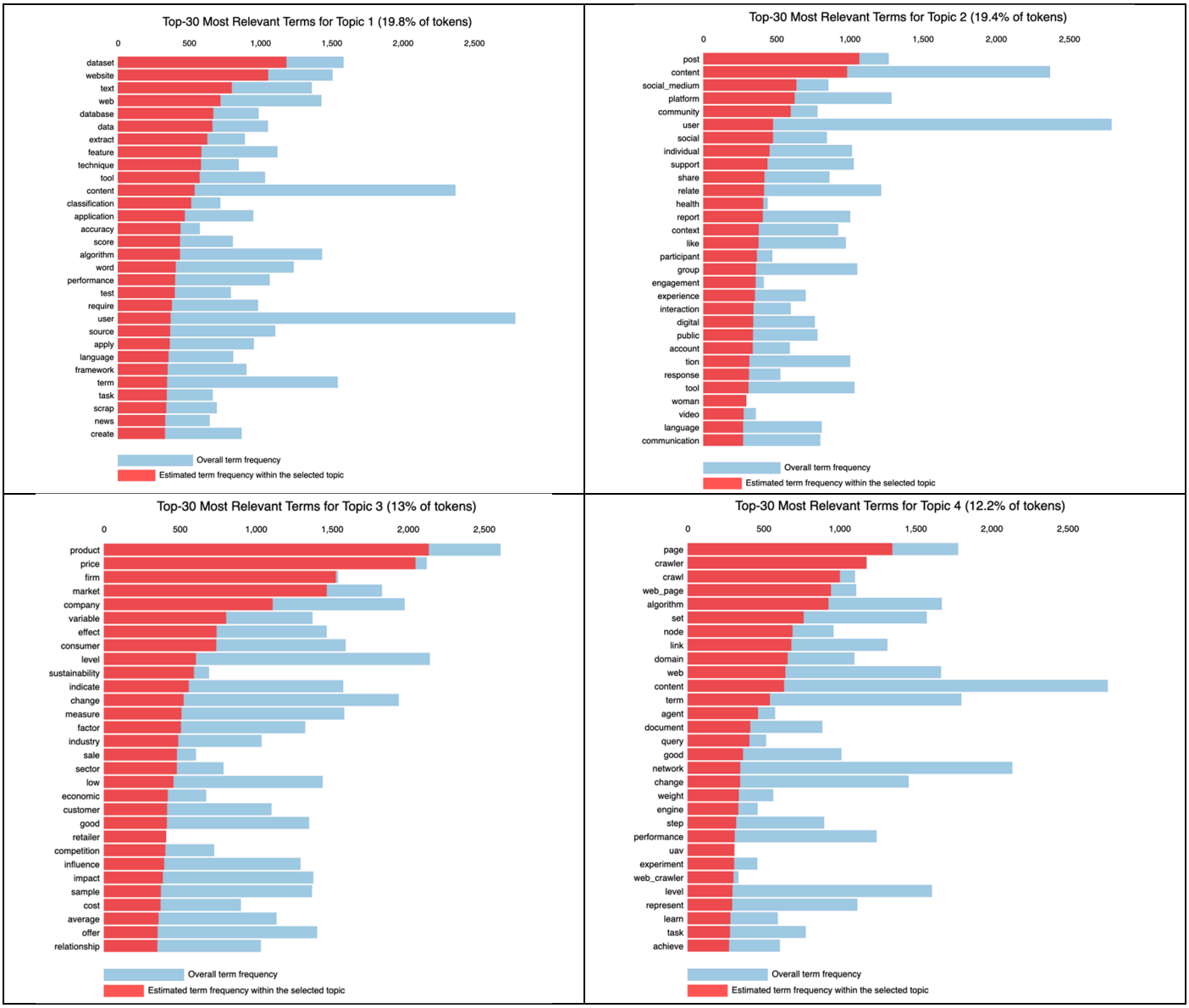
Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

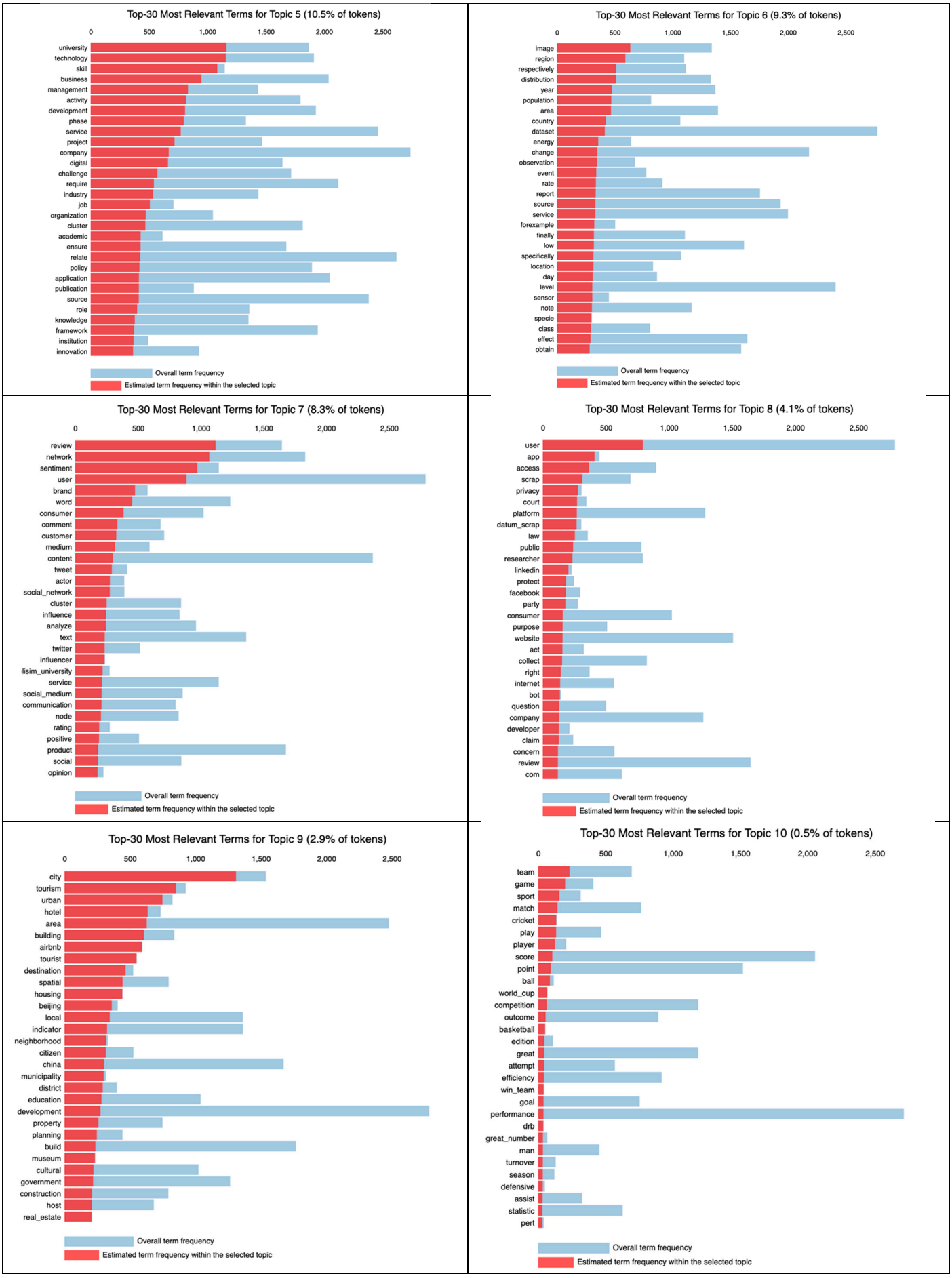
The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
IT	Information Technology
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RBV	Resource-Based View
WoS	Web of Science

Appendix A. Topic-Term Distributions



https://doi.org/10.3390/a19100868



https://doi.org/10.3390/a19100868

References

1. Baecker, J.; Weking, J.; Hein, A.; Krcmar, H. Organizational data strategy: Unveiling key elements and strategic types. *J. Inf. Technol.* **2025**, *40*, 337–361.
2. de Haan, E.; Padigar, M.; El Kihal, S.; Kübler, R.; Wieringa, J.E. Unstructured data research in business: Toward a structured approach. *J. Bus. Res.* **2024**, *177*, 114655.
3. Barba, G.; Lezzi, M.; Lazoi, M.; Corallo, A. Combined use of web scraping and AI-based models for business applications: Research evolution and future trends. *Manag. Rev. Q.* **2025**, 1–49. https://doi.org/10.1007/s11301-025-00551-3.
4. Brown, M.A.; Gruen, A.; Maldoff, G.; Messing, S.; Sanderson, Z.; Zimmer, M. Web scraping for research: Legal, ethical, institutional, and scientific considerations. *Big Data Soc.* **2025**, *12*, 20539517251381686.
5. Ashouri, S.; Hajikhani, A.; Suominen, A.; Pukelis, L.; Cunningham, S.W. Measuring digitalization at scale using web scraped data. *Technol. Forecast. Soc. Change* **2024**, *207*, 123618.
6. Conboy, K.; Mikalef, P.; Dennehy, D.; Krogstie, J. Using business analytics to enhance dynamic capabilities in operations research: A case analysis and research agenda. *Eur. J. Oper. Res.* **2020**, *281*, 656–672.
7. Mikalef, P.; Gupta, M. Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Inf. Manag.* **2021**, *58*, 103434.
8. Munir, S.; Abdul Rasid, S.Z.; Aamir, M.; Jamil, F.; Ahmed, I. Big data analytics capabilities and innovation effect of dynamic capabilities, organizational culture and role of management accountants. *Foresight* **2023**, *25*, 41–66.
9. Maluleka, M.L.; Chummun, B.Z. Competitive intelligence and strategy implementation: Critical examination of present literature review. *S. Afr. J. Inf. Manag.* **2023**, *25*, 1610.
10. Donthu, N.; Kumar, S.; Mukherjee, D.; Pandey, N.; Lim, W.M. How to conduct a bibliometric analysis: An overview and guidelines. *J. Bus. Res.* **2021**, *133*, 285–296.
11. Barbera, G.; Araujo, L.; Fernandes, S. The value of web data scraping: An application to TripAdvisor. *Big Data Cogn. Comput.* **2023**, *7*, 121.
12. Khder, M.A. Web scraping or web crawling: State of art, techniques, approaches and application. *Int. J. Adv. Soft Comput. Appl.* **2021**, *13*, 145–168.
13. Zhao, B. Web scraping. In *Encyclopedia of Big Data*; Springer International Publishing: Cham, Switzerland, 2022; pp. 951–953.
14. Bhardwaj, B.; Ahmed, S.I.; Jaiharie, J.; Dadhich, R.S.; Ganesan, M. Web scraping using summarization and named entity recognition (NER). In *Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*; IEEE: New York, NY, USA, 2021; Volume 1, pp. 261–265.
15. Dawson, N.; Rizoiu, M.A.; Johnston, B.; Williams, M.A. Predicting skill shortages in labor markets: A machine learning approach. In *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*; IEEE: New York, NY, USA, 2020; pp. 3052–3061.
16. Jorge, O.; Pons, A.; Rius, J.; Vintró, C.; Mateo, J.; Vilaplana, J. Increasing online shop revenues with web scraping: A case study for the wine sector. *Br. Food J.* **2020**, *122*, 3383–3401.
17. Wang, P.; Shea, R.; Wang, J.; Wu, E. Progressive deep web crawling through keyword queries for data enrichment. In *Proceedings of the 2019 International Conference on Management of Data*; Association for Computing Machinery (ACM): New York, NY, USA, 2019; pp. 229–246.
18. Krotov, V.; Johnson, L. Big web data: Challenges related to data, technology, legality, and ethics. *Bus. Horiz.* **2023**, *66*, 481–491.
19. Fei, L. A comparative study on public interest considerations in data scraping dispute. *Int. J. Law Context* **2024**, *20*, 568–580.
20. Fontana, A.G. Web scraping: Jurisprudence and legal doctrines. *J. World Intellect. Prop.* **2025**, *28*, 197–212.
21. Amthauer, J.; Fleiß, J.; Guggi, F.; Robertson, V.H. Detecting resale price maintenance for competition law purposes: Proof-of-concept study using web scraped data. *Comput. Law Secur. Rev.* **2023**, *51*, 105901.
22. Guyt, J.Y.; Datta, H.; Boegershausen, J. Unlocking the potential of web data for retailing research. *J. Retail.* **2024**, *100*, 130–147.
23. Macias, P.; Stelmasiak, D.; Szafranek, K. Nowcasting food inflation with a massive amount of online prices. *Int. J. Forecast.* **2023**, *39*, 809–826.
24. Muñoz-Villamizar, A.; Piatti, M.; Mejia-Argueta, C.; Pirabe, L.F.; Namdar, J.; Gomez, J.F. Navigating retail inflation in Brazil: A machine learning and web scraping approach to the basic food basket. *J. Retail. Consum. Serv.* **2024**, *79*, 103875.
25. Campos Macias, N.; Düggelin, W.; Ruf, Y.; Hanne, T. Building a technology recommender system using web crawling and natural language processing technology. *Algorithms* **2022**, *15*, 272.
26. Hartmann, P.M.; Zaki, M.; Feldmann, N.; Neely, A. Capturing value from big data—A taxonomy of data-driven business models used by start-up firms. *Int. J. Oper. Prod. Manag.* **2016**, *36*, 1382–1406.

27. Günther, W.A.; Mehrizi, M.H.R.; Huysman, M.; Deken, F.; Feldberg, F. Resourcing with data: Unpacking the process of creating data-driven value propositions. *J. Strateg. Inf. Syst.* **2022**, *31*, 101744.
28. Zeng, J.; Glaister, K.W. Value creation from big data: Looking inside the black box. *Strateg. Organ.* **2018**, *16*, 105–140.
29. Piccoli, G.; Grover, V.; Rodriguez, J. Digital transformation requires digital resource primacy: Clarification and future research directions. *J. Strateg. Inf. Syst.* **2024**, *33*, 101835.
30. Yoshikuni, A.C.; Dwivedi, R.; Zhou, D.; Wamba, S.F. Big data and business analytics enabled innovation and dynamic capabilities in organizations: Developing and validating scale. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100206.
31. Engelmann, A. A performative perspective on sensing, seizing, and transforming in small- and medium-sized enterprises. *Entrep. Reg. Dev.* **2024**, *36*, 632–658.
32. Teece, D.J. *Dynamic Capabilities: Foundational Concepts*; Cambridge University Press: Cambridge, UK, 2025.
33. Zehir, C.; Ekiz Yılmaz, T.; Kurt, A.; Borodin, A. The evolution of competitive strategy: An unsupervised machine learning approach using topic modeling and keyword clustering. *Algorithms* **2026**, *19*, 583.
34. Semerikov, S.O.; Mintii, I.S. Multi-dimensional bibliometric framework for analyzing research adaptation during crisis: Evidence from Ukrainian educational psychology. *Soc. Sci. Humanit. Open* **2026**, *13*, 102709.
35. Liu, L.; Cui, L.; Han, Q.; Zhang, C. The impact of digital capabilities and dynamic capabilities on business model innovation: The moderating effect of organizational inertia. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 420.
36. Huynh, M.T.; Nippa, M.; Aichner, T. Big data analytics capabilities: Patchwork or progress? A systematic review of the status quo and implications for future research. *Technol. Forecast. Soc. Change* **2023**, *197*, 122884.
37. Yoshikuni, A.C.; Dwivedi, R.; Kamal, M.M.; Zhou, D.; Dwivedi, P.; Apolinário, S. A dynamic information technology capability model for fostering innovation in digital transformation. *J. Innov. Knowl.* **2024**, *9*, 100589.
38. Çallı, L.; Çallı, F. Understanding airline passengers during COVID-19 outbreak to improve service quality: Topic modeling approach to complaints with latent Dirichlet allocation algorithm. *Transp. Res. Rec.* **2023**, *2677*, 656–673.
39. Ray, P.; Chakrabarti, A. A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. *Appl. Comput. Inform.* **2022**, *18*, 163–178.
40. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
41. Ogunleye, B.; Lancho Barrantes, B.S.; Zakariyyah, K.I. Topic modelling through the bibliometrics lens and its technique. *Artif. Intell. Rev.* **2025**, *58*, 74.
42. Papadia, G.; Pacella, M.; Giliberti, V. Topic modeling for automatic analysis of natural language: A case study in an Italian customer support center. *Algorithms* **2022**, *15*, 204.
43. Kastrati, Z.; Arifaj, B.; Lubishtani, A.; Gashi, F.; Nishliu, E. Aspect-based opinion mining of students' reviews on online courses. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*; Association for Computing Machinery (ACM): New York, NY, USA, 2020; pp. 510–514.
44. Cirstea, I.; Radu, A.F.; Tit, D.M.; Radu, A.; Bungau, G.; Negru, P.A. Bibliometric analysis of medical waste research using Python-driven algorithm. *Algorithms* **2025**, *18*, 312.
45. Bungau, T.; Bungau, C.C.; Bendea, C.; Hanga-Farcas, I.F.; Bendea, G. Dynamic evolution of energy efficiency in the building sector: A changepoint detection and text processing-based bibliometric analysis. *Algorithms* **2025**, *18*, 745.
46. Röder, M.; Both, A.; Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*; Association for Computing Machinery (ACM): New York, NY, USA, 2015; pp. 399–408.
47. Yao, K.; Oliva, G.A.; Hassan, A.E.; Asaduzzaman, M.; Malton, A.J.; Walenstein, A. Finding associations between natural and computer languages: A case-study of bilingual LDA applied to the Bleeping Computer forum posts. *J. Syst. Softw.* **2023**, *201*, 111651.
48. Kim, S.M.; Ha, Y.G. Automated discovery of small business domain knowledge using web crawling and data mining. In *Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp), Hong Kong, China, 18–20 January 2016*; IEEE: New York, NY, USA, 2016; pp. 481–484.
49. Li, X.; Duan, B. Organizational microblogging for event marketing: A new approach to creative placemaking. *Int. J. Urban Sci.* **2018**, *22*, 59–79.
50. Castro, L.; Debón, A.; Domenech, J. Study of the relationship between competitiveness and digital footprint indicators in Valencian wineries. In *Proceedings of the 5th International Conference on Advanced Research Methods and Analytics (CARMA 2023), Valencia, Spain, 28–30 June 2023*; Editorial Universitat Politècnica de València: Valencia, Spain, 2023; pp. 227–230.

51. Yu, C.H.; Lin, S.J. Web crawling and filtering for on-line auctions from a social network perspective. *Inf. Syst. E-Bus. Manag.* **2012**, *10*, 201–218.
52. Pinski, M.; Hofmann, T.; Benlian, A. Executive AI literacy: A text-mining approach to understand existing and demanded AI skills of leaders in unicorn firms. In *Proceedings of the International Conference on Wirtschaftsinformatik, 2023*; Springer Nature: Cham, Switzerland, 2023; pp. 219–235.
53. Laroche, F. Goodbye monopoly: The effect of open access passenger rail competition on price and frequency in France on the high-speed Paris–Lyon line. *Transp. Policy* **2024**, *147*, 12–21.
54. Chong, A.Y.L.; Ch'ng, E.; Liu, M.J.; Li, B. Predicting consumer product demands via big data: The roles of online promotional marketing and online reviews. *Int. J. Prod. Res.* **2017**, *55*, 5142–5156.
55. Parks, A.M. Unfair collection: Reclaiming control of publicly available personal information from data scrapers. *Mich. Law Rev.* **2022**, *120*, 913–945.
56. Jiang, C.; Hui, X.; Yu, J.; Huang, S.; Song, D. Configuring placemaking strategies shaping electronic word of mouth in Shanghai Shikumen heritage districts. *npj Herit. Sci.* **2025**, *13*, 153.
57. Grassetti, L.; Bellio, R.; Di Gaspero, L.; Fonseca, G.; Vidoni, P. An extended regularized adjusted plus-minus analysis for lineup management in basketball using play-by-play data. *IMA J. Manag. Math.* **2021**, *32*, 385–409.
58. Teece, D.J. Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. *Strateg. Manag. J.* **2007**, *28*, 1319–1350.
59. Day, G.S. The capabilities of market-driven organizations. *J. Mark.* **1994**, *58*, 37–52.
60. Weber, P.; Chapman, D. Location intelligence: An innovative approach to business location decision-making. *Trans. GIS* **2011**, *15*, 309–328.
61. Miah, S.J.; Vu, H.Q.; Gammack, J.; McGrath, M. A big data analytics method for tourist behaviour analysis. *Inf. Manag.* **2017**, *54*, 771–785.
62. Van den Berg, L.; Coetzee, B.; Mearns, M. Establishing competitive intelligence process elements in sport performance analysis and coaching: A comparative systematic literature review. *Int. J. Inf. Manag.* **2020**, *52*, 102071.
63. Delcea, C.; Nica, I.; Ionescu, Ș.; Cibu, B.; Țibrea, H. Mapping the frontier: A bibliometric analysis of artificial intelligence applications in local and regional studies. *Algorithms* **2024**, *17*, 418.
64. Ionescu, Ș.; Delcea, C.; Chiriță, N.; Nica, I. Exploring the use of artificial intelligence in agent-based modeling applications: A bibliometric study. *Algorithms* **2024**, *17*, 21.
65. Allaham, M.V. Bibliometric analysis of HR analytics literature. *Elektron. Sos. Bilim. Derg.* **2022**, *21*, 1147–1169.
66. Surwase, G.; Sagar, A.; Kademani, B.S.; Bhanumurthy, K. Co-citation analysis: An overview. In *Proceedings of Beyond Librarianship: Creativity, Innovation and Discovery, Mumbai, India*; B. R. Publishing Corporation: New Delhi, India, 2011; p. 9.
67. Brin, S.; Page, L. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **1998**, *30*, 107–117.
68. Heydon, A.; Najork, M. Mercator: A scalable, extensible web crawler. *World Wide Web* **1999**, *2*, 219–229.
69. Chakrabarti, S.; Van den Berg, M.; Dom, B. Focused crawling: A new approach to topic-specific Web resource discovery. *Comput. Netw.* **1999**, *31*, 1623–1640.
70. Olston, C.; Najork, M. Web crawling. *Found. Trends Inf. Retr.* **2010**, *4*, 175–246.
71. Cho, J.; Garcia-Molina, H.; Page, L. Efficient crawling through URL ordering. *Comput. Netw. ISDN Syst.* **1998**, *30*, 161–172.
72. Adar, E.; Teevan, J.; Dumais, S.T.; Elsas, J.L. The web changes everything: Understanding the dynamics of web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*; Association for Computing Machinery (ACM): New York, NY, USA, 2009; pp. 282–291.
73. Cho, J.; Garcia-Molina, H. Estimating frequency of change. *ACM Trans. Internet Technol.* **2003**, *3*, 256–290.
74. Diligenti, M.; Coetzee, F.; Lawrence, S.; Giles, C.L.; Gori, M. Focused crawling using context graphs. In *Proceedings of VLDB*; Association for Computing Machinery (ACM): New York, NY, USA, 2000; pp. 527–534.
75. Pinkerton, B. Finding what people want: Experiences with the WebCrawler. In *Proceedings of the 2nd International World Wide Web Conference*; Association for Computing Machinery (ACM): New York, NY, USA, 1994.
76. Boldi, P.; Codenotti, B.; Santini, M.; Vigna, S. UbiCrawler: A scalable fully distributed web crawler. *Softw. Pract. Exp.* **2004**, *34*, 711–726.

77. Trezza, D. To scrape or not to scrape, this is dilemma: The post-API scenario and implications on digital research. *Front. Sociol.* **2023**, *8*, 1145038.
78. Papadia, G.; Pacella, M.; Perrone, M.; Giliberti, V. A comparison of different topic modeling methods through a real case study of Italian customer care. *Algorithms* **2023**, *16*, 94.

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.