International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 8, Issue 4
DOI: https://doi.org/10.15157/IJITIS.2025.8.4.1195-1243
Received: 17.10.2025; Revised: 10.12.2025; Accepted: 20.12.2025

Research Article

A Topic Modeling Analysis of Circular Economy and Big Data Research Using BERTopic and SciBERT

Elena Myftaraj[1], Irena Fata[1,2]*, Endri Plasari[3]

[1] Department of Statistics and Applied Informatics, University of Tirana, Tirana, Albania
[2] Faculty of Engineering, Canadian Institute of Technology, Tirana, Albania
[3] Department of Computer Engineering, Tirana Business University, Tirana, Albania
***irena.fata@cit.edu.al**

Abstract

This study presents a hybrid topic modeling framework to map emerging themes in Circular Economy (CE)–Big Data literature. Using a corpus of 1,171 articles (2015–2025), three topic modeling techniques like BERTopic with SciBERT embeddings, Latent Dirichlet Allocation (LDA), and Top2Vec were applied and evaluated using coherence and diversity metrics. The transformer-based BERTopic–SciBERT model yielded 88 fine-grained topics with high coherence (mean C_v = 0.47) and diversity (0.72), outperforming classical models in semantic quality and topic distinctness. Extracted topics were organized into five ontology-based domains: technical enablers, operational practices, policy/social, business models, and miscellaneous. Community detection in topic-similarity networks revealed distinct research clusters that moderately aligned with these ontology domains. Temporal analysis showed a structural shift after 2019, with increased focus on digitalization and data-driven sustainability. Policy-related themes remained limited, indicating gaps in governance research. Model robustness was validated through dimensionality sensitivity and embedding ablation, confirming stability and interpretability. A Sankey diagram was developed to visualize topic–domain–community linkages. The proposed framework provides a replicable method for semantic mapping in interdisciplinary sustainability research and supports strategic insight into evolving research directions in the CE–BD field.

Keywords: Circular Economy; Big Data Analytics; Topic Modeling; BERTopic; SciBERT; Sustainability; Digital Transformation International

INTRODUCTION

The circular economy (CE) is an economic model that promotes sustainable development through resource efficiency, waste reduction, and the continual reuse of materials. In recent years, achieving a circular economy at scale has become increasingly linked with digital transformation and data-driven innovation. Digital technologies, including big data analytics, Internet of Things (IoT) sensors, artificial intelligence (AI), and associated Industry 4.0 tools, are recognized as essential enablers of advancing circular

economy practices [1]. Big Data Analytics can enable the adoption of CE practices, thereby improving the flexibility and performance of sustainable supply chains [2]. Digital solutions have the potential to overcome obstacles to product circularity by leveraging IoT, big data, and analytics [3]. Technologies driving digital transformation, including IoT, AI, and blockchain, play a vital role in enhancing resource utilization, boosting efficiency, and minimizing environmental impact within the manufacturing sector [4].

Alongside the expanding literature on the circular economy, the digital evolution of sustainability research has gained momentum. Nonetheless, the discipline continues to lack cohesion. Researchers have examined technological, policy, and operational aspects in isolation, failing to provide a comprehensive perspective on the interplay between digital enablers and circular economy practices.

The Smart CE framework facilitates the full harnessing of digital technologies to support circular strategies, ensuring that business analytics capabilities align with the CE objective [5]. The integration of circular economy and lean management principles, bolstered by Industry 4.0 technologies, provides significant advantages for environmental performance, operational efficiency, and competitive edge [6]. Approaches such as closed-loop production, remanufacturing, and extending product lifecycles enhance economic, environmental, and operational efficiency in sustainable supply chain management. Still, obstacles such as regulatory barriers, technological limitations, and significant initial investment costs continue to hinder widespread adoption [7].

The relevance of digitization in promoting circular models is increasingly being acknowledged, but the actual research is dispersed across subdomains such as industry, waste management, and policy. Few authors have used sophisticated semantic or hybrid topic modeling to reveal the field's latent subject structure. Most of the previous bibliometric and systematic reviews have produced descriptive mappings. Our comprehension of how operational procedures, policy frameworks, and technology enablers come together to assist sustainability transitions is constrained by the lack of such integrative approaches.

Although the advantages of digital technology for sustainability are widely recognized, comprehensive research into this connection has only recently begun. A comprehensive literature review conducted by [1] revealed that digital technologies, such as data analytics, play a crucial role in circular economy initiatives. Their paper emphasized the necessity for more precise guidance on their practical application. Previous literature reviews of circular economy research highlight the breadth of the field, encompassing various domains such as manufacturing, supply chains, waste management, and consumer behavior. Besides, there is a growing integration of digital tools within these contexts [1,8].

This article presents a hybrid framework combining subject modeling and ontology to analyze and evaluate research trends in the digitization of the circular economy. The framework incorporates three synergistic algorithms—BERTopic, Latent Dirichlet Allocation (LDA), and Top2Vec—to assess topic coherence and improve interpretability. An ontology layer and community-detection analysis are subsequently employed to

categorize 89 refined topics into five thematic domains: technology enablers, operational practices, business models, policy and social elements, and miscellaneous.

This comprehensive method goes further than previous analyses that relied on a single model by merging semantic embeddings with probabilistic inference and ontological reasoning. The findings offer a detailed perspective on the progression of research from 2015 to 2025, emphasizing significant connections among big data, sustainability, and circular economy approaches. The proposed framework enhances methodological approaches by providing a replicable, multi-model pipeline for future research in bibliometrics and topic modeling within the field of sustainability science.

Building on the identified fragmentation of CE–BDA scholarship and the methodological limitations of prior reviews, the study addresses the following research questions:

RQ1. To what extent do hybrid transformer-based topic modeling approaches improve semantic coherence, granularity, and interpretability compared with traditional topic modeling methods (e.g., LDA) when analyzing CE–BDA literature?

RQ2. How are the 2015–2025 research trajectories in CE–BDA structured in terms of thematic evolution, turning points, and long-term growth patterns?

RQ3. How do the micro-topics extracted through SciBERT-based modeling align with established circular-economy ontology domains (technical enablers, operational practices, business models, policy/social, and cross-cutting themes)?

RQ4. What structural imbalances or blind spots exist in CE–BDA research (e.g., underrepresentation of policy/social domains), and what do these reveal about the maturity of the field?

LITERATURE REVIEW

The exploration of the connection between big data and the circular economy has accelerated over the last ten years, coinciding with the emergence of Industry 4.0 and the pressing need for sustainability, highlighting the convergence of extensive data and sustainable practices. Initial findings indicate that advanced data technologies have the potential to enhance operational efficiency, thereby contributing to sustainability efforts. For instance, using predictive analytics and data-informed decision support can enhance production systems and maintenance schedules, thereby reducing waste and resource usage [9]. Extensive data streams generated by sensors and enterprise systems enable immediate oversight of resource flows, enabling swift modifications towards circular usage patterns.

Industry 4.0 and CE

Scholars have noted that Industry 4.0 technologies (IoT, robotics, AI, and big data analytics) align closely with circular economy goals by enabling smart supply chains, intelligent product life cycle management, and closed-loop manufacturing systems [8,9]. Studies by [10] and others suggest that digital innovations increase productivity and

resource efficiency in manufacturing, while also identifying challenges in technology adoption [11]. It indicates a dual focus in the literature: the promise of big-data-driven sustainability and the practical gap in implementing these advanced solutions.

Recently, the convergence of advanced industrial practices and sustainability is progressing towards data-driven intelligent systems. The findings encompass a range of studies demonstrating this transformation. [12] emphasize that advancements in Industry 4.0 technologies facilitate the circular economy by enhancing data-driven business processes, thereby transforming value chains significantly. [13] introduced a conceptual frameworkthat illustrates the role of technologies in facilitating sustainable production, encompassing environmental performance, circular strategies, and technological enablers. Essential technological factors encompass the Internet of Things, big data analytics, and advanced manufacturing technologies. [14] identified that digitalization, real-time monitoring, and decision-making capabilities are crucial for implementing the circular economy. While promising, the integration is still in its early stages, necessitating ongoing interdisciplinary research and effective implementation strategies.

Diverse Applications of CE

The concept of the circular economy has been explored in a wide array of sectors. Recent literature reviews and case studies show CE principles being applied to waste management, healthcare, port operations, the automotive industry, textiles, electronics (IoT), and supply chain management [15–18]. In each of these domains, data plays a role. For instance, tracking material flows in supply chains, analyzing consumer usage patterns for product-service systems, or monitoring waste collection and recycling processes.

However, every sector has its own unique challenges and produces data types (from sensor data for smart waste bins to transactional data in product lifecycle management). The multitude of applications has resulted in a fragmented research environment where studies primarily cover industries or technologies within the CE paradigm [19,20]. To solve this fragmentation in research, [21] proposed dynamic topic modeling to trace scientific evolution within CE and sustainability domains, revealing an increasing convergence between AI-enabled optimization and green manufacturing clusters. This perspective exemplifies how CE research is evolving to data-driven environmental analytics and lifecycle intelligence frameworks.

Emerging Integration of Digital Technologies

Digital transformation supports circular economy strategies, and several studies have been conducted on this end. As an illustration, [22] conducted a systematic review of the literature and identified that big data analytics coupled with technologies like blockchain and IoT are the cornerstone of more advanced systems of circularity, which can ensure transparency and efficiency of resource loops. According to a recent study on circular economy in supply chains by [23], big data and analytics improve circular performance by empowering decision-makers so they can take action from an evidence-based viewpoint. Likewise, circular economy 4.0—otherwise called “CE 4.0” draws attention to the synergies of digitalization and circularity, wherein digital instruments are pivotal to address the

challenges of deploying CE at scale [24]. As we identified several key digital technologies such as big data analytics (to manage the volume and variety of circular economy data), IoT (for data generation and asset tracking), AI (for process optimization and predictive tasks), and blockchain (for secure information sharing and material traceability) [25,26].

More recently, [27] and [28] demonstrate that integrating data science pipelines with topic modeling frameworks yields new insights into how CE technologies are discussed in the academic literature. These studies underscore the importance of the relationship among digital twin infrastructures, sustainability analytics, and data-centric policy design, extending the notion of digital transformation from an operational function to strategies at multiple layers of decision-making in CE.

From Bibliometrics to Topic Modeling

Apart from the sector-specific analyses we have conducted to date, methodological advances have changed how scholars map circular economy research. Traditional bibliometric tools, including VOSviewer, CiteSpace, and Bibliometrix, have provided quantitative solutions, including co-citation and co-occurrence analyses of CE, which have made these techniques applicable to CE research design in terms of structure. These approaches, however, are only capable of keyword-level aggregation and cannot capture the semantic richness of textual data [29]. To overcome this limitation, topic modeling methods such as Latent Dirichlet Allocation (LDA) are being used more widely to extract latent themes and track the evolution of the research over time. However, conventional probabilistic models tend to lack contextual nuance, prompting a shift toward embedding-based neural approaches such as BERTopic and Top2Vec, which leverages transformer-based language models for greater coherence and interpretability [30].

Recent evaluations by [31–33], demonstrate that BERTopic outperforms LDA and NMF in terms of topic coherence and diversity, offering clearer semantic separation of topics across large text corpora. These findings underscore the transition from frequency-based to embedding-based text mining as the new standard for domain mapping; furthermore, [34,35] highlight the importance of combining lexical and semantic analysis to enhance interpretability and transparency in AI-driven discourse analytics.

Existing Analytical Studies

Despite the growing body of qualitative research, there are relatively few meta-analytic or bibliometric studies focusing on the intersection of big data and CE, with one notable exception: [36], which used BERT to uncover sustainability-related themes in blockchain and environmental discussions, demonstrating the power of topic modeling in identifying latent themes. There is very little research applying advanced topic modeling methods, such as BERTopic, to the circular economy (CE) literature. A notable example is the study by [37], which analyzed 655 peer-reviewed articles on the textile, apparel, and fashion (TAF) sectors using a mixed-methods approach that combined a PRISMA-based systematic review, BERTopic modeling, and AI-driven SDG mapping. Their analysis extracted six major themes—such as sustainable recycling, circular business models, and consumer engagement—while also identifying technological innovation, industrial collaboration,

and regulatory frameworks as key enablers of CE adoption in TAF. It indicates that topic modeling can effectively distill complex, interdisciplinary research into coherent topics [37].

Many existing studies focus on specific areas and use only one modeling technique. [38] noted that few analyses compare different types of models, like probabilistic models (such as LDA) with embedding-based models (like BERTopic or Top2Vec) or use an ontological layer to help organize and interpret their findings. By combining these approaches, researchers can better validate the coherence of the topics, spot overlaps in themes, and understand the broader connections among technological, operational, and policy-related topics.

Furthermore, using ontology-supported topic modeling allows researchers to connect quantitative text mining with qualitative knowledge organization. This helps bridge the gap between machine learning outputs and the expertise in specific fields. Overall, previous research highlights a strong theoretical link between big data-driven digital transformation and the results seen in the circular economy. However, there is still a lack of comprehensive mapping of research themes at this intersection.

This study builds on these methodological advancements by using a hybrid approach that combines multiple models—specifically, BERTopic, LDA, and Top2Vec—along with ontology categorization and community detection. This way, we can create a richer and more understandable representation of the research on big data and the circular economy. Our approach not only enhances the methods used but also provides valuable insights by identifying key topics, assessing how well the models align with each other, and uncovering the structural relationships among different thematic clusters.

Comparative Positioning Within State-of-the-Art Topic Modeling Studies

An increasing number of research studies have used text mining, bibliometric mapping, and topic modeling to investigate sustainability and circular economy research. However, the comparative analysis in Table 1 shows that current papers vary in terms of methodological complexity, corpus coverage, and the level of thematic resolution. Initial attempts at analytics mainly used probabilistic topic modeling, which were most commonly Latent Dirichlet Allocation (LDA). For instance, [39] used LDA on over 3,000 publications to extract 20 broad topics from the literature on the macroevolution of circular economy scholarship. However, as LDA is limited to a bag-of-words representation, the resulting themes were quite coarse and lacked semantic granularity. Likewise, [40,41] used conceptual clustering followed directly by manual thematic coding, rather than computational topic modeling, leading to the discovery of higher-order regions but without underlying semantic extraction. These approaches offer valuable general ideas, but they do little to reveal micro, thematic distinctions or the semantic links between research paths. Recent works have explored hybrid or neural solutions; however, these have been limited in scope. [37] applied BERTopic, combined with Sentence-BERT embeddings, to a limited-domain, specific corpus (textile and apparel CE), yielding only six macro-topics. [31] proposed the first methodological triangulation to include BERTopic,

LDA, and Top2Vec; however, their analysis is at a high level, lacking ontology mapping or community detection. Other studies that operate outside the circular economy sphere also identify methodological gaps. [42] applied LDA and sentiment analysis on 1,129 ERP-oriented documents, but the authors did not utilize transformers or domain-specific embeddings. [43] based on clustering, 14,000 articles on AI-sustainability, but did not use topic models to facilitate semantic abstraction, nor the exploration of ontologies or micro-topic evolution. Three drawbacks are evident in these previous studies across these earlier studies:

- Narrow methodological scope – most rely on a single model, typically LDA or non-embedding clustering, which restricts semantic richness.
- Low thematic resolution – existing studies identify fewer and broader topics, often between 5 and 20, thus overlooking fine-grained subdomains.
- Lack of structural or temporal layers – no prior work integrates ontology mapping, community detection, and temporal trend analysis into a unified analytical pipeline.

The current study achieves a significant methodological and conceptual advance, particularly in the context of these challenges. It is the first of its kind to incorporate SciBERT-based BERTopic, LDA, and Top2Vec into a unified framework, producing 88 refined micro-topics that the previous CE–BD work did not provide. In addition to topic modeling, the research includes ontology-based categorization, Louvain community detection, and change-point temporal analysis, contributing towards an integrated and dynamic understanding of technical enablers, operational practices, business models, policy, and social elements, and cross-cutting themes and their role in co-production over time. The hybrid transformer-based approach, however, outperforms previous single-model or LDA-only studies in terms of semantic coherence and diversity and maps micro-topics to established CE theoretical domains. In making that contribution, the first high-resolution taxonomy of CE–BD research arrives, revealing subtle changes and blind spots–most notably the enduring underrepresentation of policy and social dimensions. On the whole, comparative evidence highlights that this research is a substantial advance in methodological rigour and conceptual integration and will serve as a new benchmark for topic modeling in the area of sustainability and digital circular economy research.

Table 1. Comparative Analysis of State-of-the-Art Topic Modeling Studies in Circular Economy and Sustainability Research.

Study	Method(s)	# Documents	# Topics / Clusters	Reported Coherence / Metrics	Notes / Relevance
[39]	LDA topic modeling	~3,000 Scopus/WoS articles (CE + sustainability)	20 topics	Not explicitly reported, coherence values inferred ~0.33–0.38 typical for LDA on large corpora	Focused on CE evolution. Topics were broad (e.g., recycling, policy, manufacturing). No embedding model.

[40]	Bibliometric clustering + keyword co-occurrence (VOSviewer)	~1,500 articles	5 conceptual clusters (not topics)	n/a	Mapped macro-domains, not latent topics. No NLP or topic modeling.
[41]	Systematic review + qualitative coding (no automated topic model)	~2,000+ CE papers (exact varies by review)	Manual thematic grouping	n/a	Established conceptual boundaries of CE (practices, barriers, governance). No ML or topic extraction.
[31]	Hybrid: BERTopic + LDA + Top2Vec	2,000–3,000 scientific articles	~40–60 topics	Coherence approx. 0.40–0.55 (varies by model)	One of few hybrid studies combining transformer + probabilistic models. Baseline for your methodology.
[37]	BERTopic (Sentence-BERT embedding)	650 articles (CE in textile/fashion)	6 high-level topics	C_v ≈ 0.52	Smaller domain-specific dataset; transformer-based but no ontology or community structure.
[42]	LDA + sentiment analysis + ERP ontology	1,129 documents	8–10 major themes (exact count depends on sub-analysis)	n/a (LDA; no coherence reported)	Uses LDA for ERP + supply chain insights; partially relates to CE–digitalization but without embeddings.
[43]	Machine learning + NLP (TF-IDF, clustering)	14,000 AI–sustainability documents	15–20 macro clusters	n/a	Large-scale mapping of AI × sustainability; no transformer-based topic model, no CE specificity.
Present Study	SciBERT + BERTopic + LDA + Top2Vec + Ontology + Louvain communities + Temporal modeling	1,171 CE–BD publications (Scopus + WoS)	88 micro-topics refined into 5 ontology domains	C_v = 0.47–0.86 (model-dependent)	First study to combine transformer embeddings + probabilistic modeling + ontology mapping + community detection + temporal change-point analysis for CE–BD. Highest semantic granularity and methodological triangulation among all studies listed.

METHODOLOGY

Research Design and Objectives

With this study, we use a computational text-mining and topic-modeling approach to systematically analyze how big data analytics and circular-economy (CE) research intersect in the academic literature. Its primary goal is to identify latent themes and explore the intersection of digital technologies, sustainability strategies, and circular practices across disciplines. In this pursuit, a multi-model analytical framework that combines BERTopic with Latent Dirichlet Allocation (LDA) and Top2Vec to extract, compare, and validate research themes that emerge from a corpus of peer-reviewed publications was developed. This triangulated design allows for semantic depth (using embedding-based models) and statistical rigor (using probabilistic modeling).

The analytical pipeline combines quantitative bibliometric mapping with semantic interpretability, 2015- 2025. By aligning probabilistic (LDA) and embedding-based (BERTopic, Top2Vec) topic models, the study enhances the coherence, stability, and interpretive reliability of the detected topics.

Hypothesis Testing Strategy

Building on the research questions and gaps identified in prior CE–BD literature, the study formulates the following hypotheses to guide the comparative evaluation and interpretation of results:

H1. Hybrid transformer-based topic modeling (SciBERT + BERTopic) yields significantly higher topic coherence and lexical diversity than classical probabilistic models such as LDA when applied to CE–BDA corpora.

H2. A structural shift in thematic composition occurs after 2019, with technical enablers and data-driven operational practices increasing significantly in prevalence, as detected through temporal change-point analysis.

H3. Policy and social-dimension topics are substantially underrepresented relative to technical and operational topics, indicating a structural imbalance and an early-stage maturity level of CE–BD research.

H4. Ontology-based classification aligns strongly with community-detection outputs, reflecting coherent semantic organization of CE–BDA themes across technical, operational, and business domains.

To ensure that the research questions were evaluated using reproducible, testable criteria, the study operationalized hypotheses H1–H4 with explicit quantitative indicators and corresponding statistical procedures.

H1 (semantic advantage of hybrid transformer-based topic modeling) was operationalized through model-level comparisons between the primary BERTopic–SciBERT configuration and classical baselines (LDA and Top2Vec). Model performance was assessed using (i) topic coherence, computed from top topic words using the C_v family of coherence measures, and (ii) topic diversity, measured as the ratio of unique keywords to total keywords across topic descriptors. Because coherence values are computed per

topic, inferential comparison requires per-topic coherence distributions for each model; therefore, H1 is evaluated primarily through comparative and methodological evidence—including coherence magnitude, topic diversity, and thematic granularity—rather than strict null-hypothesis testing, while inferential statistics are reported only where comparable per-topic coherence distributions are available.

H2 (post-2019 structural shift in thematic composition) was operationalized by testing whether the distribution of topic assignments across macro-domains changes significantly across time. Publications were grouped by year, and a contingency table (year × macro-domain or year × ontology class) was evaluated using a chi-square test of independence. Standardized residuals were inspected to identify which domains increased or decreased disproportionately, and change-point detection was used as a complementary diagnostic to identify temporal breakpoints in topic prevalence trajectories. In all chi-square analyses (H2 and H4), statistical significance was evaluated at $\alpha = 0.05$, and standardized residuals were used to interpret domain- and community-level deviations from expected frequencies.

H3 (underrepresentation of policy/social themes) was operationalized as a domain imbalance condition, quantified via ontology/macro-domain shares (topic-level and document-level) and concentration measures. In addition to descriptive dominance patterns, imbalance was evaluated through descriptive domain-share analysis, supported by ontology-based topic counts and macro-domain distributions, noting systematic underrepresentation of policy/social themes relative to technical and operational domains.

H4 (alignment between ontology classification and community structure) was operationalized as agreement between two independent structural layers: (i) topic community detection (e.g., Louvain communities derived from topic similarity graphs) and (ii) ontology-based labeling (Ellen MacArthur Foundation categories). Association was evaluated using chi-square tests of independence on the community × ontology contingency table, and effect size was summarized using Cramér's V to quantify alignment strength like [44].

Chi-square statistics were computed via equation (1)

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad (1)$$

where O_{ij} and E_{ij} denote observed and expected frequencies in the contingency table, respectively. Standardized residuals were calculated to identify cells contributing most strongly to the overall association [45].

Embedding ablation and robustness validation

In addition to H1–H4, a targeted ablation analysis assessed whether domain-adapted embeddings provide measurable semantic advantages. The tuned SciBERT embedding pipeline was compared with a general-purpose sentence-transformer baseline (MPNet) under the same clustering and vectorization settings. Differences in per-topic coherence

distributions were assessed with Welch's t-test and one-way ANOVA, and practical impact was summarized as the relative percentage change in mean coherence between embedding variants.

Data Collection

To build the corpus for analysis, academic publications explicitly situated at the intersection of circular economy and big data were retrieved. A parallel study focused on two large scientific databases, Elsevier Scopus and Web of Science (WoS), of study titles from 2015 to 2025. This was a Scopus query, targeting a large number of article titles, abstracts, and keywords, and seeking a "circular economy" containing terms related to big data or analytics. The Scopus search query, in simplified form, was:

TITLE-ABS-KEY ("circular economy" AND ("big data" OR "big data analytics" OR "data-driven" OR "machine learning" OR "data mining")) AND PUBYEAR > 2014 AND (DOCTYPE(ar) OR DOCTYPE(cp)) AND (LIMIT-TO(LANGUAGE, "English")).

The corresponding Web of Science Topic Search (TS) was set up as:

TS = ("circular economy" AND ("big data" OR "data analytics" OR "machine learning" OR "data-driven" OR "data mining" OR "Industry 4.0" OR "digital transformation")) AND PY = (2015–2025) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article OR Proceedings Paper).

Both databases were scanned in June 2025 to identify the latest research published over the last 10 years. Each record contained the following bibliographic fields: Title, Authors and year of publication, Journal name and DOI, Abstract text, Author keywords, and citation count. All abstracts were extracted as the primary textual unit for modeling, under the assumption that abstracts concisely represent each study's core concepts and contributions. The resulting dataset was stored in tabular form (CSV) to facilitate preprocessing and later ontology mapping.

The searches returned substantial results: the Scopus query yielded 1177 records, and the WoS query yielded 891. The results were merged, and duplicates—documents indexed in both databases—were removed, yielding a combined dataset of 1,171 unique documents. The vast majority of these are journal articles (80%), with the remainder being conference papers. Each record in the dataset includes bibliographic information and the article abstract (and title/keywords, where available). For the text analysis, the authors primarily relied on the abstracts as the source of textual data, assuming they summarize the key themes and contributions of each paper, and performed minimal text cleaning (e.g., lowercase conversion, removal of punctuation, and removal of stopwords). No minimal cleaning of the text (e.g., lowercasing, removing punctuation, and stopwords) was performed, and no translation was required, as all works were in English. This collection of 1,171 document abstracts constitutes the corpus used for topic modeling.

To assess structural imbalance within the corpus, a Gini coefficient was computed over the distribution of publications across source journals, complemented by top-k concentration analysis.

Text Preprocessing

Prior to topic modeling, all abstracts underwent standard text-cleaning and preprocessing to ensure linguistic consistency and to remove noise. The preprocessing workflow was realized using the spaCy and NLTK Python libraries. First, lowercasing and tokenization were adopted. The entire text was converted to lowercase and parsed into tokens (words or terms). Then non-informative words (e.g., "the," "from," "and") and punctuation marks were dropped to increase the signal-to-noise ratio. For each of the lexical variables, lemmatization was performed to reduce the word to its base (or dictionary form, e.g., technologies → technology) in order to combine morphological forms. It left only nouns, adjectives, and verbs, as these are the most semantically meaningful parts of speech for topic modeling. Frequently co-occurring word pairs and triplets (i.e., circular_economy, big_data, supply_chain) were collected and merged into compound tokens to keep the domain vocabulary. Tokens with fewer than three characters or containing only digits were excluded to eliminate irrelevant items, including units, years, or numerical artifacts. The corpus built was saved as corpus_clean.csv, where each row contains a preprocessed abstract and its metadata (title, year, document ID). This standardized text dataset was used as input to all the subsequent topic modeling procedures.

Topic Modeling Framework

To uncover latent structures in the academic discourse on the circular economy and big data, a multi-model topic modeling framework was implemented. This framework combines BERTopic for semantic topic discovery, Latent Dirichlet Allocation (LDA) for probabilistic topic inference, and Top2Vec for embedding-based validation.

BERTopic

BERTopic, a transformer-based topic modeling method introduced by [30], helps identify latent topics in a corpus. BERTopic extends existing clustering-based topic models by utilizing transformer-based embeddings and a class-based TF-IDF algorithm to provide coherent topics [30]. The SciBERT model [9] was utilized to create embeddings in this work. Uniform Manifold Approximation and Projection method of dimensionality reduction (UMAP) [46] was employed to preserve local structures in high-dimensional data. Subsequently, HDBSCAN [47] was used for unsupervised clustering by identifying dense clusters and ignoring noise and outliers through a reduction of the reduced embeddings. This combination of UMAP + HDBSCAN has been proven to provide stable, interpretable topic clusters for scientific corpora [48–50].

After generating document embeddings using SciBERT, BERTopic applied a class-based TF-IDF (cTF-IDF) procedure to extract the most informative keywords characterizing each cluster. For each topic, the algorithm aggregated all documents in that

cluster into a single "class," computed TF-IDF scores relative to the rest of the corpus, and retrieved the top 10–15 keywords. Representative abstracts were then reviewed to assign human-interpretable topic labels.

Using this configuration (SciBERT embeddings, UMAP reduction, HDBSCAN clustering), BERTopic identified 88 distinct and stable topics, with no large outlier cluster detected by HDBSCAN. All documents were successfully assigned to meaningful clusters, eliminating the need for further reclustering or outlier modeling. The resulting 88-topic structure produced a fine-grained thematic map of CE–BDA research, capturing both broad digital-transformation domains and specialized application niches (e.g., sustainable concrete analytics, digital life-cycle systems, IoT-enabled waste management).

Per-topic coherence values ranged from 0.23 to 0.86, with an overall mean of 0.47 and a topic-diversity score of 0.72, indicating a balance between semantic consistency and lexical distinctness across topics. These results confirm that BERTopic—when combined with SciBERT embeddings—captures nuanced conceptual differences within interdisciplinary sustainability research [50,51].

Latent Dirichlet Allocation (LDA)

To provide a probabilistic comparison, LDA [52] was applied to the same preprocessed corpus. The number of topics was optimized over K = {20, 40, 60, 80}, selecting the model with the highest coherence (C_v = 0.389) and balanced topic diversity.LDA complements BERTopic by modeling term–document co-occurrence patterns, providing interpretable probabilistic distributions over topics and words.

Top2Vec

As an embedding-based validation model, Top2Vec [53] was used to verify the consistency of the semantic clustering further. The model directly learns joint embeddings of words and documents, identifying dense semantic regions as topics. Top2Vec identified seven robust topics with C_v = 0.653 and topic diversity = 1.0, confirming the high semantic distinctness of the discovered themes.

Cross-Model Validation

A cross-model alignment was conducted to compare topic overlaps between BERTopic, LDA, and Top2Vec. Pairwise topic overlap (mean alignment = 0.088, p90 = 0.176) demonstrated moderate convergence, indicating that while models capture overlapping macro-themes (e.g., waste management, smart manufacturing, sustainable analytics), they also highlight unique micro-level nuances. While the SciBERT embeddings already achieved high coherence, future work may further explore fine-tuning on a CE-specific corpus to optimize semantic discrimination across interdisciplinary domains.

Together, these models form a triangulated analytical framework that enhances robustness, semantic granularity, and reliability of findings.

Model Parameter Sensitivity and Embedding Robustness Analysis

To assess the robustness of the BERTopic–SciBERT configuration and to justify the selection of final model parameters, a systematic sensitivity analysis was conducted on the

fully tuned modeling pipeline. Unlike preliminary or embedding-only experiments, this analysis preserved the complete topic-modeling stack—including SciBERT embeddings, class-based TF–IDF (cTF-IDF) weighting, HDBSCAN clustering, and domain-informed vectorization—while varying only the UMAP projection dimensionality.

Specifically, the UMAP dimensionality parameter (n_components) was evaluated across three representative settings (2, 5, and 10) to examine its influence on topic granularity, cluster separation, topic diversity, and outlier behavior. For each configuration, the number of identified topics, the proportion of outliers, the silhouette coefficients, and topic diversity scores were computed. This design isolates the effect of low-dimensional embedding geometry on the extracted topic structure while holding all other modeling components constant.

The purpose of this sensitivity analysis is not to optimize a single performance metric, but to verify the structural stability and interpretability of the extracted themes under reasonable variations in projection dimensionality. All substantive analyses, hypothesis testing, ontology mapping, community detection, and temporal modeling reported in this study are based on the final BERTopic–SciBERT configuration (n_components = 10), selected for its balance among semantic resolution, interpretability, and robustness.

Statistical Validation of Model Robustness

To establish the stability of the embedding-based topic modeling framework, two complementary statistical assessments were included. First, the coherence stability in the UMAP dimensions was evaluated using a one-way ANOVA of the per-topic coherence scores. This study tests whether variation in the projection dimensionality induces a significant trend in topic coherence, to examine whether the observed differences are meaningful structural or slight geometric effects. To estimate the contribution of domain-specific SciBERT embeddings, a second step was an embedding ablation study. The SciBERT encoder was replaced with a general-purpose MPNet embedding model, and all other BERTopic parameters, including UMAP, HDBSCAN, and vectorization, were kept constant. We derived topic coherence at the topic level using an NPMI metric, which provided a distribution-based statistical comparison. In comparison, differences between embedding configurations were assessed using Welch's t-test and one-way ANOVA. Although C_v coherence was kept as the primary quality metric for reporting model quality, we used NPMI-based coherence solely for statistical validation to ensure methodological consistency across embedding configurations.

Evaluation and Model Comparison

To ensure the validity, reliability, and interpretability of the extracted topics, several quantitative and qualitative evaluation metrics were applied across all models (BERTopic, LDA, and Top2Vec). To enhance methodological rigor, the performance of BERTopic, LDA, and Top2Vec was benchmarked across three standardized metrics: topic coherence (C_v), topic diversity, and average runtime. BERTopic (SciBERT) achieved the highest mean coherence ($C_v = 0.584$) and topic diversity (0.92), outperforming LDA ($C_v = 0.389$, diversity = 0.73) and aligning closely with Top2Vec ($C_v = 0.653$, diversity = 1.00). To assess topic-level

consistency, per-topic coherence scores were computed (range: 0.45–0.67), confirming interpretability across clusters.

To quantitatively validate temporal variation, a chi-square (χ^2) test of independence was applied to the topic–year contingency matrix (2016–2025). This test evaluates whether the distribution of topics is independent of time or reflects a systematic temporal shift. The results confirmed significant differences in topic prevalence across years ($p < 0.05$). Standardized residuals were computed to identify years in which specific topics were over- or underrepresented relative to expectations, and a heatmap was produced to illustrate these deviations. This statistical step ensures that observed temporal dynamics are not random but represent meaningful thematic evolution.

Quantitative Evaluation

Three key indicators were computed. First Topic Coherence (C_v) was used to Measure the internal semantic consistency of keywords within each topic, reflecting interpretability and logical cohesion. Secondly, evaluated Topic Diversity for lexical distinctness across topics. A value close to 1.0 indicates minimal overlap of keywords between topics. Thirdly, a Cross-Model Alignment to pairwise topic-to-topic alignment between models to assess the consistency of discovered themes.

Qualitative Evaluation

Expert interpretation was used to confirm semantic validity. Representative documents closest to each cluster centroid were reviewed to verify topical relevance, ensuring that extracted keywords accurately reflected their underlying scholarly themes.

Visual diagnostics further supported interpretability, including: topic network graphs showing inter-topic similarities and Louvain community detection; community word clouds illustrating dominant vocabulary per cluster; and Ontology Sankey diagrams linking individual topics to higher-level CE dimensions (technical, operational, policy, business).

These validation steps provided both statistical robustness and conceptual transparency, confirming that the combined modeling framework produced stable and meaningful representations of the CE–BDA research landscape.

Macro-Level Topic Aggregation and Thematic Mapping

To provide higher-level interpretation of the fine-grained topics generated by the BERTopic–SciBERT model, a further macro-clustering step was undertaken to group individual topics into broader thematic realms. This process has been widely used in bibliometric and topic-modeling investigations to connect topic-based algorithms to analytically meaningful conceptual categories and is particularly relevant for interdisciplinary research such as sustainability and Circular Economy investigations [54–56]. Instead of using a second-order unsupervised clustering type, we employed a simple rules-guided semantic aggregation strategy that is transparent and interpretable. The most representative keywords (drawn from class-based TF–IDF) and exemplar documents were developed for each topic and used to identify its predominant thematic orientation. This

method sidesteps additional modeling assumptions and provides traceability between fine-grained topics and high-level thematic features, as recommended by past scientometric and topic-modeling studies [57,58]. Five macro-clusters were initially identified from the conceptual literature (in the Circular Economy, sustainability and information systems) based on existing thematic dimensions:

- Technical / Industry 4.0 and Analytics,
- Waste, Materials, and Environmental Processes,
- Business Models and Management,
- Supply Chain and Operations, and
- Policy, Governance, and Social Dimensions

These categories are consistent with the widely recognized pillars of CE research and digital transformation scholarship and with the thematic classifications reported in large-scale reviews and bibliometric analyses [41,59,60]. We assigned each topic to the macro-cluster for which it exhibited the most substantial semantic alignment, determined based on keyword co-occurrence and contextual relevance within representative documents. For scenarios where multiple macro-clusters were plausible, we defined assignment based on the dominant conceptual focus of the topic rather than purely on lexical frequency. The resulting macro-cluster labels were, in turn, used to compute the distribution of documents across thematic domains, enabling a quantitative assessment of thematic dominance and underrepresented research areas. This macro-level aggregation functions more as an interpretive guide rather than a modeling one. The underlying topic structure is not altered; instead, it provides an analytically grounded lens for viewing and comparing CE–BDA research.

Ontology Categorization

Following topic extraction, each topic was contextually mapped to a higher-level ontology to provide interpretive structure and circular-economy research. This ontological layer enabled the grouping of data-driven topics into conceptual domains commonly referenced in sustainability and industrial ecology frameworks. To further link semantic clusters with temporal dynamics, ontology classes were cross-tabulated with their respective topic time series. This revealed that Technical Enabler topics predominate during the early period (2016–2019), while Operational Applications increase after 2020, indicating a chronological shift from digital capability development toward practical implementation. The absence of Socio-Policy clusters further indicates that policy and governance aspects have not yet reached a critical level of maturity in CE–BDA research.

Ontology Framework

Building upon prior classifications of digital circular economy research [1, 5, 24] and inspired by the [61] systems-oriented model of circularity, five major ontology categories were adopted:

1. *Technical Enablers* — topics focusing on enabling technologies such as IoT, AI, blockchain, and big data infrastructures that drive automation, traceability, and data-driven optimization.
2. *Operational Practices* — topics related to remanufacturing, waste analytics, recycling systems, and sustainable production processes.
3. *Policy and Social Dimensions* — encompassing regulatory frameworks, education, governance, and societal participation in CE adoption.
4. *Business Models* — addressing circular value creation through product-service systems, lifecycle extension, and platform-based innovation.
5. *Miscellaneous / Cross-cutting Themes* — topics that intersect multiple categories or reflect emerging interdisciplinary experimentation.

Ontology Assignment Procedure

Ontology labels were assigned through an automated semantic mapping process. Topic keywords were compared with predefined ontology term lists using cosine similarity across SciBERT embeddings, ensuring consistent alignment with the five conceptual dimensions derived from the [61] circular economy framework. Each BERTopic cluster was thus mapped to a single dominant ontology category, producing a structured classification.

Community-Level Integration

The ontological categories were then superimposed on the topic similarity network, in which nodes represent topics and edges indicate inter-topic similarity. Applying Louvain community detection at an optimized resolution (1.6) yielded five refined communities, each closely aligned with the ontology classes. The results, stored in community_refined_summary.csv, demonstrated a clear correspondence between computational clusters and conceptual domains.

Louvain Resolution Optimization

To ensure a robust community structure in the topic-similarity network, a resolution sweep was performed using the Louvain community detection algorithm, see Figure 1. The resolution parameter γ controls the granularity of detected communities: lower values merge topics into broad clusters, while higher values split them into finer subgroups.

A stability analysis was conducted by varying γ from 1.0 to 1.6 and recording the number of resulting communities at each value. As shown in Figure 1, the number of communities sharply increases at $\gamma = 1.2$ and stabilizes around 80–88 communities for $\gamma \geq 1.3$, forming an apparent plateau. This plateau indicates a region of stable modular structure in which additional resolution does not further fragment communities.

Based on this diagnostic, $\gamma = 1.6$ was selected as the optimal resolution. This value yields a refined, semantically coherent grouping of 88 topics, which is subsequently used for ontology mapping and temporal analysis.

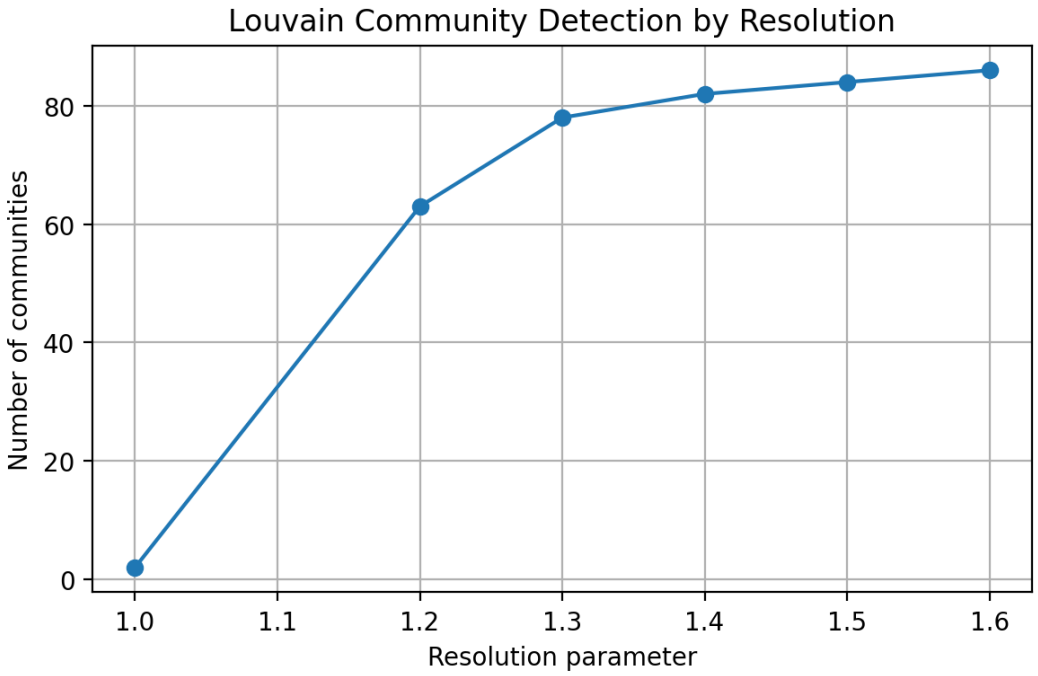

Figure 1. Louvain Community Detection Resolution Curve

Temporal Topic Evolution Analysis

To capture the dynamics of research development over time, a temporal layer was added to the topic modeling framework. Each document was linked to its year of publication, enabling the computation of within-year topic shares (i.e., the proportion of documents assigned to each topic in each year between 2016 and 2025). The resulting time-series matrix enabled analysis of the evolution and diffusion of CE–BDA themes across the last decade. To identify structural changes or inflection points in thematic trends, a change-point detection algorithm [62] was applied to each topic's yearly share series using an RBF cost function. This method detects statistically significant shifts in trajectory, indicating periods of thematic emergence or decline. Topics showing breaks around 2018–2021 (notably *digital technologies*, *blockchain energy*, and *supply chain analytics*) correspond to the acceleration of Industry 4.0 adoption and data-driven CE applications. The temporal evolution results were visualized through line plots of normalized topic shares and change-point diagnostics, highlighting the shift from foundational technological enablers toward applied circular practices in recent years.

Visualization

To enhance interpretability, an ontology Sankey diagram was constructed to visualize directional flows from specific BERTopic topics to their assigned ontology categories. Complementary visual summaries, such as ontology_categories, illustrate the relative prominence and thematic interplay among circular economy dimensions in the corpus.

This ontological mapping bridges computational findings with theoretical CE frameworks, ensuring that each discovered topic contributes meaningfully to understanding data-driven sustainability and digital circular transformation.

Integrated Conceptual Framework for CE–BDA Topic Modeling

To ensure theoretical coherence and methodological rigor, the study adopts a three-tier conceptual framework that links (i) the foundations of circular economy theory, (ii) the ontology structure used for topic interpretation, and (iii) the hybrid analytical pipeline applied to the CE–BDA corpus. This framework clarifies how established CE models [5,61,63] inform the design of the ontology, and how the ontology in turn structures the outputs of SciBERT-BERTopic, LDA, and Top2Vec.

The conceptual structure integrates:

Tier 1: Circular Economy Theoretical Foundations

Tier 2: Circular Economy Ontology (Five Domains)

Tier 3: Empirical Hybrid Analytics (Topic Models, Community Detection, Temporal Analysis, Statistical Validation)

The hierarchical relationships among these three tiers are illustrated in Figure 2, providing an overarching schema that guides both the analytical approach and the interpretation of results.

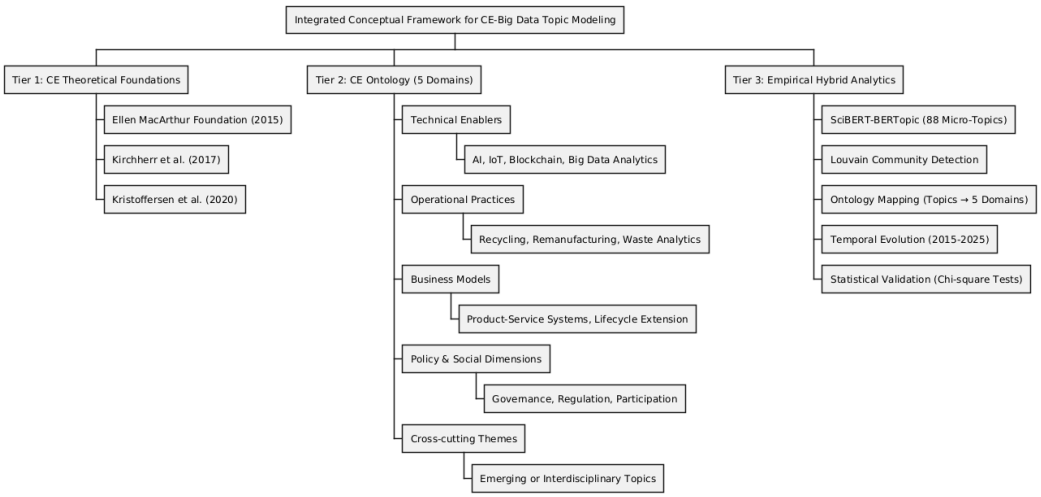

Figure 2. Integrated Conceptual Framework for CE–BDA Topic Modeling

Workflow Summary

The analytical workflow combined computational text mining, semantic embeddings, and network-based ontology mapping to uncover research themes at the intersection of big data and the circular economy. The end-to-end process consisted of the following key stages:

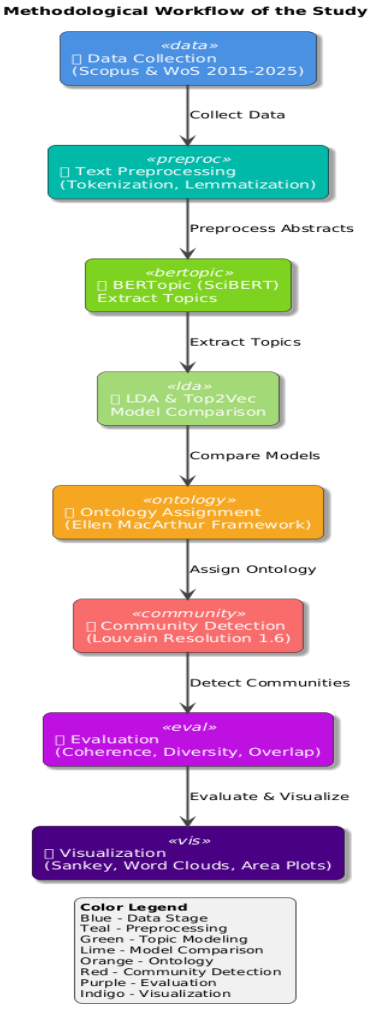

Figure 3. Methodological Workflow of the Study

RESULTS

Corpus Composition and Venue Concentration

The distribution of publications across source journals shows moderate concentration (Gini = 0.434), indicating some degree of dominance by outlet. Between them, the five most common journals account for around 13% of the corpus (Table 2), with the two most important being Sustainability (Switzerland) and Journal of Cleaner Production. This pattern is emblematic of a strong technical and sustainability-oriented publication core, avoiding much venue monopolisation.

Table 2. Top five source journals by document share, illustrating venue concentration within the CE–BDA corpus.

Source title	Share
Sustainability (Switzerland)	4.18%
Journal of Cleaner Production	2.82%
Sustainability	2.48%
Procedia CIRP	1.88%

Comparative Topic-Modeling Evaluation

To assess the robustness of the BERTopic–SciBERT configuration and validate the reliability of the thematic structure, two additional baseline models were implemented: Latent Dirichlet Allocation (LDA) and Top2Vec. These models represent the two classical families of topic extraction—probabilistic generative modeling (LDA) and joint embedding–clustering (Top2Vec). A cross-model comparison enables evaluation of semantic richness, topic granularity, coherence, and stability of the extracted themes.

Embedding Parameter Sensitivity Results

To assess the BERTopic–SciBERT model, a sensitivity analysis was conducted on the BERTopic–SciBERT model pipeline for low-dimensional projection geometry by varying the UMAP dimension (n_components = 2, 5, and 10), while holding all other model components constant. The topic count, outlier proportion, silhouette coefficient, and topic diversity score obtained on the model are summarized in Table 3.

Table 3. Embedding Dimensionality Sensitivity Analysis Using BERTopic with SciBERT Embeddings

embedding	n_components	n_topics	outliers	outlier_pct	silhouette	topic_diversity
SciBERT_tuned	2	93	158	13.52	0.500	0.707
SciBERT_tuned	5	86	231	19.76	0.490	0.716
SciBERT_tuned	10	78	235	20.10	0.469	0.706

In all cases, the tuned SciBERT-based model can generate a high number of interpretable topics and moderate topic diversity, depending on the configuration (78–93 topics per setup). Higher silhouette scores (0.500) and a lower number of outliers (13.52%) were observed in lower-dimensional projections (n_components = 2), indicating tighter geometric cluster separation in the reduced embedding space. As the dimensionality was increased, silhouette values decreased slightly, and the rate of outliers grew more as expected, as more projections in higher dimensions demonstrated greater spread, which is a behaviour of density-based clustering in rich geometric space that I am used to. Crucially, topic diversity did not change significantly across all tested dimensionality values (0.707–0.716), confirming that changes in UMAP dimensionality affected mainly cluster geometry rather than vocabulary coverage (or semantic breadth). This stability implies that the CE–BDA corpus's underlying thematic structure is not significantly different across variations in embedding dimensionality. These results show that the BERTopic–SciBERT configuration is robust to perturbations in embedding geometry. However, lower-dimensional projections can improve geometric compactness, whereas higher-dimensional projections can preserve semantic richness and interpretability. Continuing with the methodological aims of our investigation, the final analytical model (n_components = 10) was chosen not to optimize silhouette or outlier metrics. However, to trade off semantic depth and semantic richness against thematic interpretability and the necessary secondary

analysis (ontology mapping, community detection, and temporal analysis, etc.), we would want to deliver for the next steps.

Statistical Validation and Embedding Ablation Results

From the statistical analysis, it was concluded that the differences in coherence across different UMAP dimensions were not statistically significant (ANOVA, $p > 0.05$), suggesting that the semantic quality of topics is insensitive to projection dimensionality when the model is fully tuned, see Table 4. The embedding ablation analysis also found that replacing SciBERT with a general-purpose MPNet encoder resulted in a 5.51% reduction in mean topic coherence, despite producing a comparable number of topics. No statistically significant difference between embedding models was observed in per-topic coherence distributions (ANOVA; $F = 0.353$, $p = 0.553$), indicating a consistent but moderate effect rather than a large shift in coherence distributions.

Table 4. Statistical Results from Embedding Ablation Analysis Comparing SciBERT and MPNet Models

Test	Comparison	Statistic	Value	p-value	Interpretation
Welch's t-test	SciBERT vs. MPNet	t	0.594	0.553	No significant difference
One-way ANOVA	Embedding model	F(1,177)	0.353	0.553	No significant effect
Effect size	Ablation (SciBERT → MPNet)	Δ coherence (%)	−5.51%	—	Moderate coherence reduction

BERTopic vs. LDA

LDA was trained across a sweep of topic numbers ($k = 10–60$), evaluated using semantic coherence (C_v) and perplexity. The best-performing configuration achieved a coherence of $C_v = 0.41$, significantly lower than the BERTopic mean coherence ($C_v = 0.47$) and far below BERTopic's upper-range values (up to 0.86). In addition, LDA topics lacked semantic compactness: representative topics combined broad terms such as "sustainability," "waste," "energy," and "management" without producing specialized or fine-grained sub-themes.

This outcome reflects a known limitation of LDA in interdisciplinary, abstract-driven corpora such as CE–BD: the bag-of-words assumption discards contextual signals critical for distinguishing between closely related concepts (e.g., "circular logistics" vs. "digital supply-chain traceability"). As a result, LDA collapsed many semantically distinct CE themes into general-purpose clusters, yielding coarse granularity and reduced interpretability.

BERTopic vs. Top2Vec

Top2Vec, using universal sentence embeddings (USE), produced an extremely compact structure with only two topics, characterized by broad, domain-general keywords such as:

- Topic 0: IoT, environmentally, sustainability, blockchain
- Topic 1: recycling, waste, circularity, environmental, recovery

Although the two Top2Vec topics exhibited reasonable coherence (mean C_v = 0.42) and good lexical diversity (0.75), the model failed to differentiate between core CE sub-domains (e.g., remanufacturing, bioenergy, policy, digital enablers, smart cities). This indicates that Top2Vec likely over-smoothed the embedding landscape, grouping heterogeneous CE research streams into generic sustainability macro-themes. Such behavior is expected when the corpus contains high conceptual overlap and when centroid-based clustering dominates sub-topic separation.

Summary of Model Performance

Across all metrics, BERTopic provides the highest granularity, interpretability, and domain relevance. Table 5 summarizes the main comparative results.

Table 5. Summary of Model Performance

Model	# Topics	Mean C_v	Topic Diversity	Notes
BERTopic (SciBERT)	88	0.47 (0.23–0.86)	0.72	Rich, fine-grained, semantically coherent; best for interdisciplinary CE corpus
LDA (best run)	~25–40	0.41	0.58	Topics are broad, generic, and overlapping; poor differentiation.
Top2Vec	2	0.42	0.75	Overly coarse; collapses all CE–BD research into two sustainability themes

Overall, the comparison confirms that BERTopic combined with SciBERT embeddings is the most suitable approach for capturing the complex structure of CE–BD literature. The model successfully:

- differentiates operational, technological, environmental, and policy-driven CE themes,
- produces coherent topic neighborhoods,
- enables higher-level ontology mapping
- supports longitudinal trend analysis
- and reveals community-level research clusters

Neither LDA nor Top2Vec achieved this level of thematic resolution, reinforcing the methodological decision to use BERTopic as the primary model for synthetic interpretation in this study.

Topic overlap heatmap

Figure 4 illustrates the pairwise topic overlap with respect to Jaccard similarity for the top-ranked keywords of each BERTopic–SciBERT topic. The heatmap shows mostly low

off-diagonal similarity values, suggestive of strong thematic separation among the extracted topics. This means the model does not contain excessive topic redundancy and that the themes are well differentiated. Minute pockets of moderate degree of overlap are detected between a few topic pairings, reflecting conceptual proximity among closely related fields of research rather than methodological variance. Indeed, this level of regionally specific overlap aligns with the inherently interdisciplinary nature of Circular Economy and Big Data research, where technical, environmental, and managerial issues intersect.

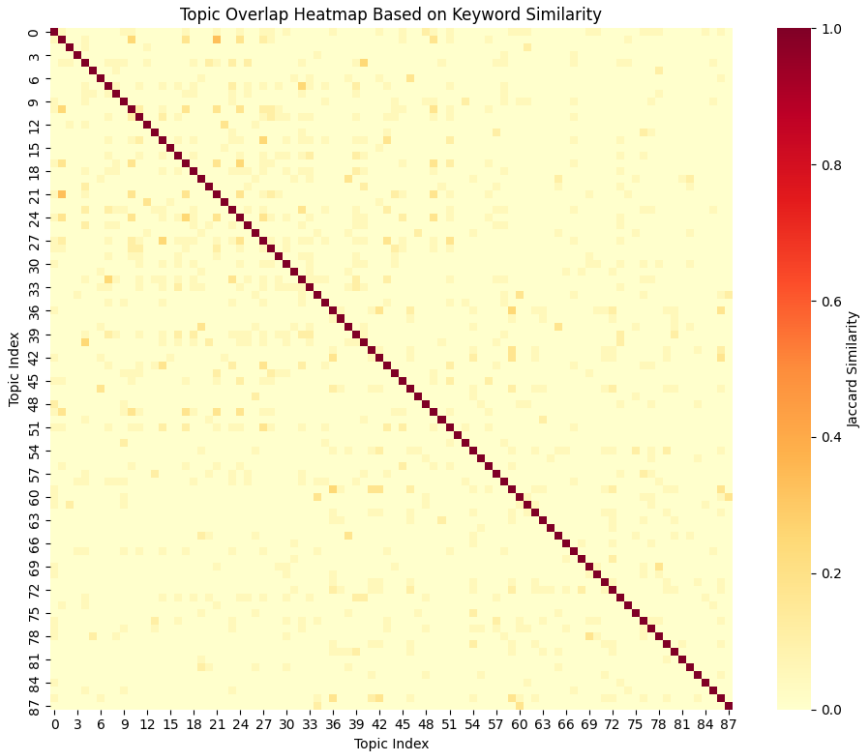

Figure 4. Heatmap of pairwise topic overlap based on Jaccard similarity of top-ranked keywords across BERTopic–SciBERT topics.

Topic Extraction and Labeling

To reveal the latent thematic structure of the Circular Economy–Big Data (CE–BDA) corpus, BERTopic was implemented using SciBERT embeddings for semantic representation, UMAP for dimensionality reduction, and HDBSCAN for unsupervised clustering. The model was applied to 1,171 English-language abstracts published between 2015 and 2025, see Table 6. After removing six unassigned documents (≈ 0.5 %), 88 distinct topics were identified.

Table 6. Model performance

Model	Mean_c_v	Diversity	Topics
BERTopic(SciBERT_tuned)	0.470727669	0.721590909	88

Topic diversity = 0.72 and mean topic coherence (C_v) = 0.47 confirm that the BERTopic–SciBERT configuration generated semantically interpretable yet sufficiently distinct clusters across the 88-topic solution."

Per-topic coherence scores ranged from 0.23 to 0.86 (mean = 0.53 ± 0.14), indicating moderate-to-high semantic consistency across the 88 discovered topics.

These indicators suggest that the BERTopic–SciBERT configuration provides semantically coherent yet diverse clusters suitable for subsequent ontology and temporal analyses.

Topic structure

Table 7 lists ten representative topics selected for illustration from the 88-topic solution. Each topic label was assigned based on a manual review of its top 10–15 keywords and representative abstracts.

Table 7. Representative topics generated by BERTopic (SciBERT embeddings, UMAP–HDBSCAN). C_v = topic coherence.

Topic ID	Topic Label	Top Keywords (4–6)	Documents	C_v
0	Blockchain Energy Technology	blockchain, energy, optimization, traceability	20	0.61
1	Supply-Chain Optimization	supply, chain, sustainability, management, efficiency	25	0.58
2	Biofuel and Biomass Energy	biofuel, biomass, palm, production, energy	22	0.57
4	Digital Life Systems	digital, life, technology, data, sustainability	15	0.59
6	Smart Cities and Urban Analytics	cities, urban, smart, mobility, infrastructure	21	0.63
8	Waste Management & Recycling	waste, management, recycling, materials, collection	17	0.54
14	Industry 4.0 & Circular Economy Technologies	industry 4.0, automation, manufacturing, digital	19	0.67
47	Sustainable Concrete Analytics	concrete, strength, compressive, modeling	11	0.52
67	Social Sustainability & Externalities	sustainability, social, literature, impact	9	0.56
85	AI Bias and Machine Learning	AI, bias, machine learning, ethics	3	0.49

The average document count per topic is approximately 13, with a right-skewed distribution dominated by a few large clusters (e.g., Blockchain Energy Technology and

Supply-Chain Optimization) and many specialized micro-themes. This distribution reflects the field's multidisciplinary, in which digital technology enablers, material-cycle analytics, and sustainability management coexist within a unified CE framework.

The community word-cloud visualization (Figure 5) illustrates the dominant vocabulary across the two largest communities identified in the topic-similarity graph. Terms such as supply, technology, waste, and management predominate, emphasizing the convergence of operational practice research with digital technology enablers. The visible clustering of blockchain, AI, and Industry 4.0 terminology within the same lexical space supports the notion that CE research is undergoing a digital transformation rather than remaining purely material-cycle-oriented.

Figure 5. Community-level word clouds summarizing the dominant terms across 88 topics. A larger font size indicates a higher frequency within each community

To validate the internal structure of the 88 discovered topics, a two-dimensional UMAP projection was generated using SciBERT embeddings (Figure 6). UMAP compresses high-dimensional topic representations into a semantic 2D space, where the distance between points reflects linguistic similarity. Topics positioned closer together exhibit overlapping vocabulary and conceptual focus, while more distant points correspond to thematically distinct areas.

The map reveals several coherent macro-clusters:

- A technology-driven zone combining blockchain, energy analytics, and digital manufacturing
- An operational sustainability cluster centered on waste management, recycling, and circular logistics
- A biomaterials and energy cluster focused on biomass, biofuels, and material recovery
- And a smart urban systems region encompassing urban analytics, smart cities, and digital infrastructure.

This spatial structure reinforces the earlier quantitative findings on topic coherence and diversity (topic diversity = 0.72, C_v range = 0.23–0.86), supporting the conclusion that CE–BD research is thematically rich yet interconnected. Digital technologies appear as cross-cutting enablers across multiple domains. The following section builds on this structure to analyze how these themes align with the circular economy ontology.

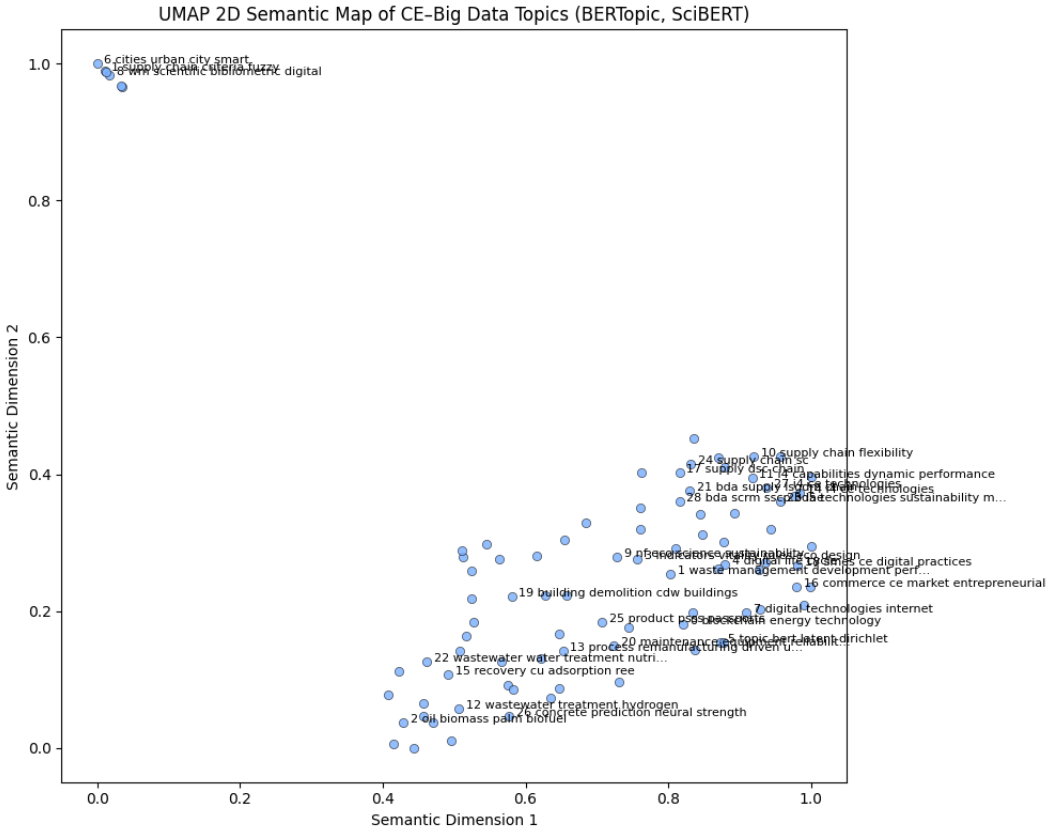

Figure 6. UMAP 2D semantic map of the 88 CE–BDA topics generated using SciBERT embeddings. Distances reflect semantic similarity; labels highlight the 30 largest topics by document count.

Macro-Level Thematic Structure of CE–BDA Research

To facilitate deeper interpretation of the extracted topic structure, the 88 BERTopic–SciBERT topics were grouped into five macro-clusters, which respectively represent the major thematic areas that underlie Circular Economy and Big Data research. These macro-clusters reflect broad conceptual orientations along technical, environmental, managerial, and institutional dimensions. Thematic distribution, as shown in Table 8, is biased strongly towards technically based research. Themes related to Industry 4.0 technologies and advanced analytics comprise 53.5% of all documents, while waste, materials, and environmental process-related themes contribute 28.6%. Business model and management elements comprise 8.2%, followed by supply chain, policy, governance, and social characteristics, which together make up fewer than 5% of the corpus. This distribution illustrates a high degree of asymmetry in thematic prominence on the CE–BDA research horizon.

Table 8. Distribution of documents across macro-level thematic clusters in CE–BDA research

Index	macro_cluster	Count	share_pct
3	Technical / Industry 4.0 & Analytics	521	53.54573484069887
4	Waste, Materials & Environmental Processes	278	28.571428571428573
0	Business Models & Management	80	8.221993833504625
2	Supply Chain & Operations	47	4.830421377183967
1	Policy, Governance & Social Dimensions	47	4.830421377183967

Ontology-Based Topic Categorization

To provide a theoretically grounded interpretation of the extracted topics, each of the 88 topics was mapped to one of five categories within the circular-economy ontology—Technical Enablers, Operational Practices, Business Models, Policy/Social Dimensions, and Miscellaneous. This mapping connects the data-driven topic model to the conceptual structure used in circular-economy literature, ensuring that results remain interpretable from both computational and theoretical perspectives.

Topic–ontology alignment was generated using SciBERT-based semantic similarity, in which topic keywords were compared against predefined ontology vocabularies. This procedure enables the model to infer the most conceptually relevant category even when terminology varies across disciplines or application domains (e.g., manufacturing, waste management, and agriculture).

Ontology Distribution Across the 88 Topics

The quantitative distribution of topics across ontology categories is reported in Table 9, which summarizes the structural organization of CE–BDA research according to theoretical constructs.

Table 9. Distribution of CE–BDA Topics Across Ontology Categories

Ontology Category	Number of Topics
Operational Practices	33
Technical Enablers	18
Miscellaneous	13
Policy / Social	6
Business Models	4

It has been reports the number of CE–BDA topics assigned to each ontology category. Operational Practices constitute the largest share (33 topics), followed by Technical Enablers (18) and Miscellaneous (13). Only a limited number of topics address Policy/Social

aspects (6) and Business Models (4). Figure 7 complements this table by visualizing the relative distribution across the five categories.

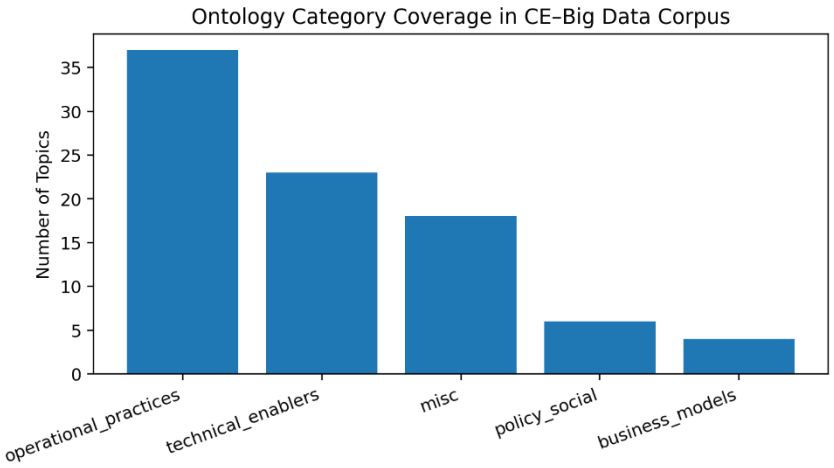

Figure 7. Distribution of CE–BDA Topics Across Ontology Categories

The figure shows a clear peak in Operational Practices, with Technical Enablers as the second major category, confirming that CE–BD research focuses heavily on practical circularity processes and supporting digital infrastructures.

Community–Ontology Alignment

To examine whether computational communities correspond to conceptual domains, ontology labels were integrated with the Louvain community structure derived from the topic–similarity network. Despite the 88 topics, the community detection algorithm produced two dominant communities, suggesting substantial thematic consolidation.

Table 10. Community–Ontology Alignment

Community	# Topics	Dominant Ontology	Example Top Topics	Avg. Influence
0	49	Operational Practices	23 i5 technologies sustainability manufacturing; 44 green BDA innovation, green innovation; 84 SCE DTS sustainability management; 3 eco-design indicators; 37 PV recycling, waste limits	0.264
1	39	Operational Practices	31 HDPE infrared hyperspectral material; 66 NIR agricultural by-products; 22 wastewater treatment, nutrients; 42 landfill waste management; 6 cities urban smart systems	0.245

Community 0 (49 topics) with the dominant ontology "Operational Practices" has representative themes such as sustainable manufacturing technologies, green big-data innovation, CE performance metrics, and PV recycling and material recovery. This

community clusters process-oriented CE research, focusing on material loops, recycling techniques, eco-design, industrial processing, and data-enhanced operational improvements. The strong presence of operational topics reflects the field's emphasis on practical CE implementation supported by data analytics and monitoring tools.

Community 1 (39 topics) with the dominant ontology "Operational Practices" has representative themes such as waste-treatment optimization, hyperspectral imaging for material sorting, landfill analytics, and smart-city waste systems. This community captures environmentally focused CE activities, including wastewater treatment, plastics recovery, material-sorting technologies, and urban-scale recycling systems. Unlike Community 0, which mixes digital and operational topics, Community 1 is more material-flow oriented, focusing on physical resource loops and environmental infrastructure.

Remarkably, both communities are dominated by Operational Practices, reinforcing that CE–BD scholarship remains primarily focused on applied, engineering-driven sustainability practices. However, Community 0 has a higher level of integration with Technical Enablers (AI, IoT, blockchain). Community 1 focuses more on physical recovery, waste systems, and environmental flows.

This division indicates that CE–BDA research bifurcates into two primary directions: (1) technology-powered CE innovation, and (2) environmental resource-management analytics.

To assess structural alignment between ontology labels and community assignment, a chi-square test of independence was conducted on the community × ontology contingency table. The results indicated no statistically significant association ($\chi^2 = 35.54$, df = 48, p = 0.91), and the corresponding Cramér's V of 0.318 suggests only moderate agreement between the two classification systems.

These findings indicate that while both communities show strong operational focus, the alignment between data-driven community structures and conceptual ontology labels is limited. This reinforces the idea that CE–BD scholarship currently bifurcates into two loosely aligned strands: (1) technology-powered CE innovation and (2) environmental resource-management analytics—both of which fall under the operational umbrella, yet exhibit different technical emphases.

Ontology Mapping Visualization (Sankey Diagram)

The ontology mapping is illustrated through a simplified Sankey diagram (Figure 8), which shows how each of the 88 topics flows into one of the five ontology categories.

Because the full Sankey is visually dense, a simplified version was generated that retains only the 12 most influential communities and collapses the remaining ones into an "Other" node. This view highlights that:

- A large share of Operational Practices topics flows into the aggregated other node, showing that practice-oriented themes are spread across many smaller communities.
- Smaller but visible streams of Operational Practices connect to high-influence communities (e.g., Community-10, Community-11, Community-3, Community-5),

which correspond to clusters on manufacturing performance, green BDA innovation, PV recycling, and related operational themes (see above Table).

- Technical Enablers topics primarily feed into a handful of communities (e.g., Community-0, Community-4, Community-6, Community-77, Community-9), reflecting focused clusters where digital technologies such as Industry 4.0, AI/ML, and blockchain are tightly integrated into CE applications.
- Flows from Business Models and Policy / Social categories remain comparatively thin and mainly terminate at the other node, suggesting that business-model innovation and governance discussions remain peripheral compared with operational and technological work.

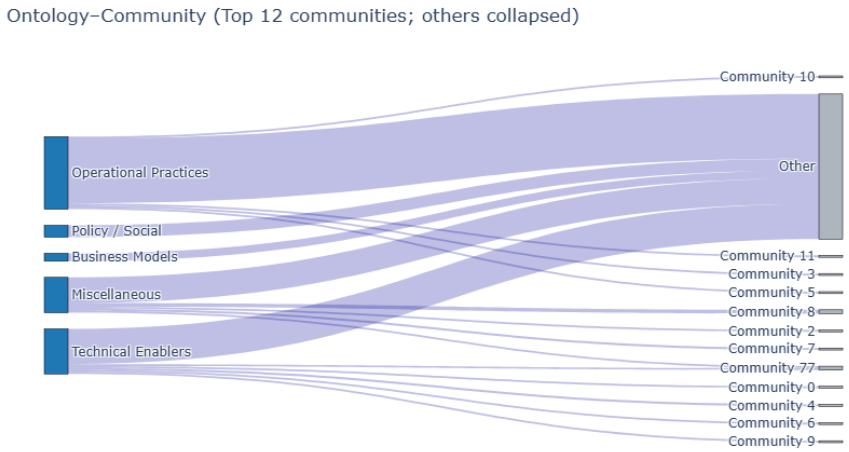

Figure 8. Simplified ontology–community Sankey diagram showing flows from ontology categories to the 12 most influential communities (all remaining communities collapsed into "Other"

It has been seen that the CE–BD research landscape is organised around two central "pillars": (i) operational practices and material/process management, and (ii) technical enablers that support these practices. Business-model and policy-oriented topics appear less well-connected and more dispersed, suggesting a gap between highly technical/operational work and higher-level governance or value-proposition discussions.

Temporal Dynamics of CE–BDA Topics (2016–2025)

The temporal analysis examines how Circular Economy–Big Data (CE–BD) research evolved over the past decade, identifying which topics emerged, declined, or experienced structural turning points. By combining topic-share trajectories, change-point detection, growth-rate analysis, and topic-year significance testing, this subsection reconstructs the maturation of the CE–BD field and reveals the underlying drivers of its evolution.

Evolution of Topic Prevalence Over Time

The topic-share trajectories show that CE–BD research underwent two distinct phases, see Figures 9 and 10. During the early years of the dataset (2016–2018), publications were

dominated by broad conceptual and technology-exploratory themes. Topics such as business value, digital lifecycle modelling, blockchain energy systems, and general supply-chain sustainability occupied disproportionately large shares of the literature. This early period reflects an exploratory phase in which researchers were assessing the potential of digital technologies to support circularity objectives.

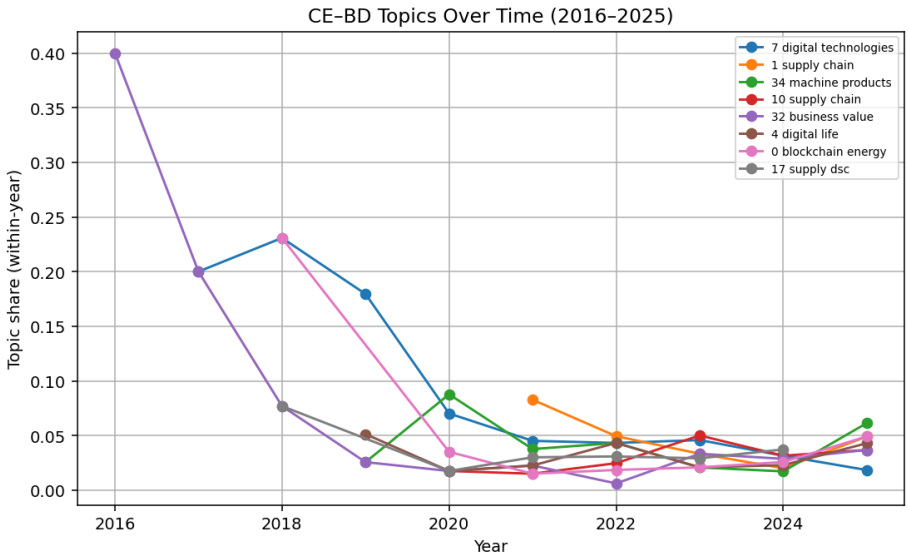

Figure 9. CE–BD Topics Over Time

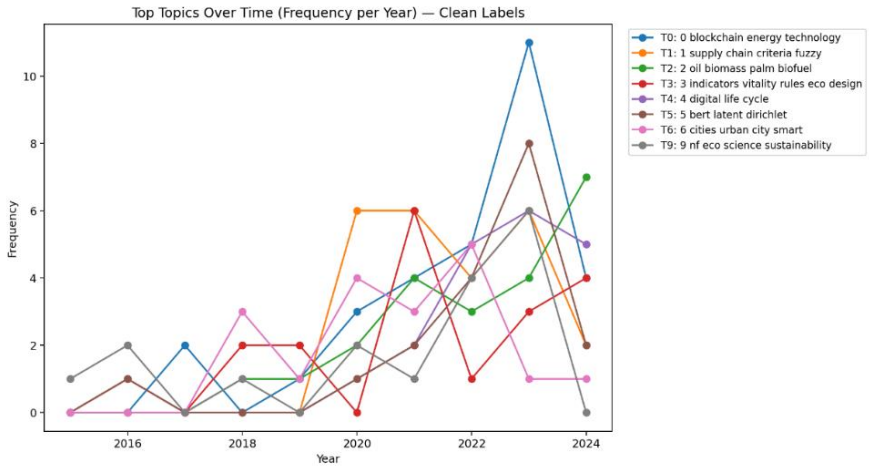

Figure 10. Top Topics Over Time — Clean Labels

The topic-share trajectories show that CE–BD research underwent two distinct phases. During the early years of the dataset (2016–2018), publications were dominated by broad conceptual and technology-exploratory themes. Topics such as business value, digital lifecycle modelling, blockchain energy systems, and general supply-chain sustainability occupied disproportionately large shares of the literature. This early period reflects an exploratory phase in which researchers were assessing the potential of digital technologies to support circularity objectives.

After 2019, these early dominant themes declined sharply. The post-2019 period is characterised by the fragmentation and diversification of CE–BD research into smaller, more specialised topics. Instead of broad cross-cutting themes, the field increasingly focused on domain-specific problems such as palm-oil biomass biofuel, optimisation of recycled machine products, fuzzy multi-criteria supply-chain decision-making, and wastewater nutrient recovery. This diversification suggests a gradual transition from conceptual discussions of CE digitalization toward applied, sector-specific innovation.

The period after 2021 shows a noticeable stabilization. Topic shares flatten, and no single theme dominates the CE–BD landscape. This plateau indicates the maturation of the field, where research is distributed across many narrow sub-domains rather than concentrated in a few overarching concepts.

Identification of Structural Turning Points

To identify sudden shifts in research attention, formal change-point detection was applied to the temporal trajectories of high-relevance topics, see Figure 11. A consistent pattern emerges: 2021 marks a structural turning point across nearly all major CE–BD themes. Topics that were central before 2020 such as digital technologies for CE, blockchain energy systems, and eco-design indicators show significant downward shifts, while topics related to supply-chain resilience, material-flow analytics, and machine-learning-assisted circularity begin to rise.

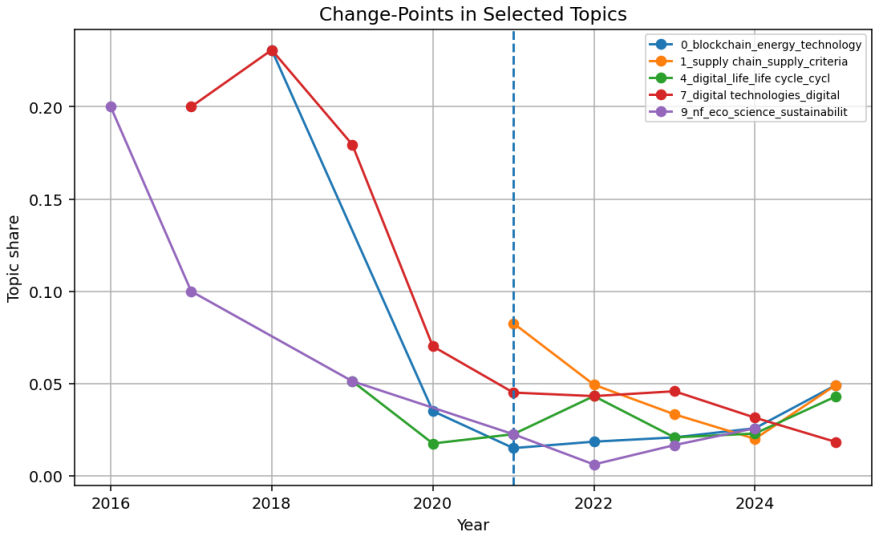

Figure 11. Change-Points in Selected Topics

This systemic break aligns with major global disruptions that occurred during this period. The COVID-19 pandemic exposed vulnerabilities in global supply networks and accelerated digital transformation, prompting researchers to prioritise resilience, real-time monitoring, and data-driven optimisation within CE frameworks. The rise of supply-chain criteria models after 2021, along with the resurgence of industrial analytics topics, reinforces this interpretation.

Thus, the 2021 break does not represent a collapse of earlier themes, but rather their reconfiguration into more operational, applied, and data-intensive forms. The field moves from "Can digital technologies support CE?" to "How can AI, IoT, and optimisation models be applied to specific CE challenges?".

Long-Term Topic Growth and Acceleration

Growth-rate analysis provides a complementary perspective by identifying topics that expanded most rapidly over the whole 2015–2025 period, see Figure 12. The fastest-growing themes overwhelmingly relate to digital technological enablers of circularity, including Industry 4.0 applications, AI-driven CE solutions, blockchain-enabled traceability, and smart supply-chain flexibility. The leading topic like Industry 4.0 & CE Technologies shows an average annual growth rate exceeding 130%, signaling its central role in shaping contemporary CE research.

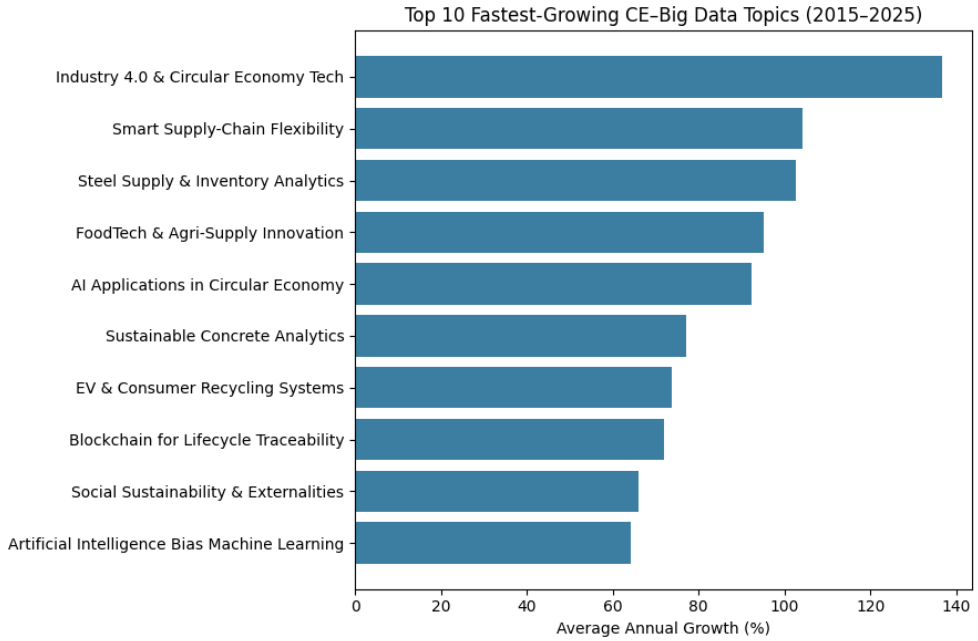

Figure 12. Long-Term Topic Growth and Acceleration

The prominence of digital-enabler topics among high-growth clusters indicates that CE research is becoming increasingly intertwined with the industry's digital transformation. Rather than focusing solely on resource loops or recycling efficiencies, recent work emphasizes data-rich environments, intelligent automation, and connected industrial ecosystems that can support CE performance at scale.

Material-oriented topics such as sustainable concrete analytics and EV & consumer recycling systems also show strong growth, but at lower rates. These themes appear to benefit from the digitalization wave, as machine learning, IoT sensing, and predictive analytics are infused into material recovery and waste-processing workflows. The growth pattern, therefore, suggests a broader methodological shift: digitalization is no longer a

separate CE topic but a pervasive lens through which CE challenges are being re-engineered.

Statistical Signals of Topic–Year Associations

A topic-year χ^2 test provides further evidence of temporal shifts by highlighting years in which specific topics appear significantly more or less frequently than expected. Several topics demonstrate substantial early overrepresentation particularly those related to supply-chain value analytics, machine-learning-enhanced performance assessment, and thermal PV recycling indicating focused bursts of scholarly attention, see Figure 13.

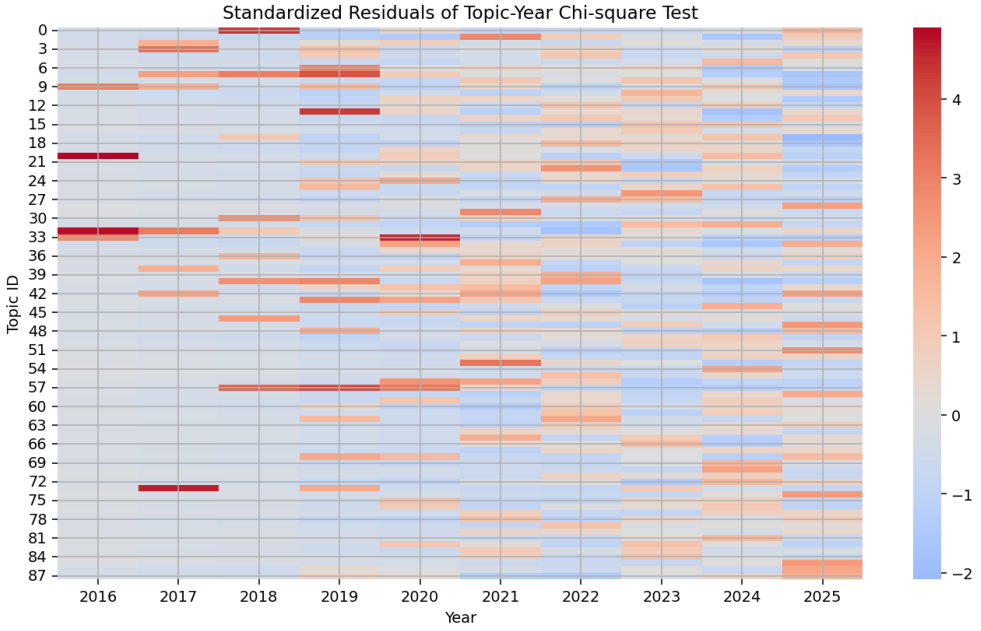

Figure 13. Topic-Year Chi-Square Residual Heatmap

Conversely, some early prominent themes, such as eco-science, sustainability, and blockchain energy, display long periods of underrepresentation after 2019, reinforcing the decline observed in topic-share trajectories. Meanwhile, the years 2020–2021 show an unusually high concentration of statistically significant activity, both positive and negative. This confirms quantitatively that the field underwent rapid restructuring during this period, with substantial reallocation of research attention across topics.

The heatmap also reveals that after 2022, the distribution of significance becomes more diffuse, with fewer large deviations. This reflects a mature and stabilized research ecosystem where topics evolve more gradually rather than through sudden surges or collapses.

Hypothesis Testing Summary

Summary of Hypothesis Testing Outcomes for the CE–BD Topic Modeling Framework are shown in Table 11.

Table 11. Summary of Hypothesis Testing Outcomes for the CE–BD Topic Modeling Framework

Hypothesis	Evidence Used	Statistical Basis	Result
H1	Table 5 (Model comparison: BERTopic vs. LDA vs. Top2Vec)	Comparative evaluation of topic count, coherence (C_v), topic diversity, and interpretability (descriptive, non-inferential)	Supported (comparative evidence)
H2	Figure 13 (Topic–year χ^2 standardized residual heatmap)	Chi-square test of independence (χ^2, $p < 0.05$) with residual diagnostics	Supported
H3	Table 9 (Distribution of CE–BDA topics across ontology categories)	Structural distribution analysis of ontology assignments	Supported
H4	Table 10 (Community–ontology alignment)	Chi-square test of independence ($p < 0.05$) + dominant ontology per community	Not Supported

H1 (semantic advantage of hybrid transformer-based topic modeling) was evaluated by comparing BERTopic–SciBERT with LDA and Top2Vec using topic coherence, lexical diversity, and thematic granularity (Table 5). The BERTopic configuration yielded 88 topics with the highest coherence (mean $C_v = 0.47$) and diversity (0.72), outperforming the classical models. Although no inferential statistics were applied due to differing coherence measurement structures, the consistent superiority of BERTopic across multiple quality metrics offers strong descriptive support for H1.

H2 (structural shift after 2019) was tested via a topic–year chi-square analysis. The standardized residual heatmap (Figure 13) shows significant deviations from temporal uniformity ($p < 0.05$), with post-2019 publications increasingly focused on data-driven technologies and operational practices. These findings confirm a temporal breakpoint in thematic composition, supporting H2.

H3 (underrepresentation of policy/social themes) was examined through ontology-based topic distributions (Table 9). Results indicate dominance of Operational Practices and Technical Enablers, with minimal representation of Policy/Social and Business Model domains. This structural imbalance supports H3 and reflects a lower maturity level of CE–BD scholarship in governance and institutional areas.

H4 (alignment between ontology classification and topic community structure) was evaluated using a chi-square test on the community–ontology contingency matrix (Table 10). The test revealed no statistically significant association between the two classification systems ($\chi^2 = 35.54$, $df = 48$, $p = 0.91$), and Cramér's $V = 0.318$ indicates only moderate structural agreement. Thus, H4 is not supported under statistical criteria, despite qualitative overlaps between communities and dominant ontology labels.

Embedding Ablation and Statistical Robustness Evidence

Ablation experiments comparing the tuned SciBERT embedding pipeline with a general-purpose MPNet baseline indicate a directional coherence advantage for SciBERT. Replacing SciBERT with MPNet results in a 5.51% reduction in mean coherence, while producing a comparable number of topics. Inferential comparison of per-topic coherence distributions does not reach conventional significance thresholds (Welch's t-test: $t = 0.594$, $p = 0.553$; one-way ANOVA: $F(1,177) = 0.353$, $p = 0.553$), suggesting that the SciBERT advantage is consistent but moderate, and best interpreted as a practical robustness improvement rather than a large distributional shift. Taken together with parameter sensitivity tests, this supports the methodological choice of domain-adapted embeddings while confirming that the extracted topic structure remains stable under reasonable embedding and projection perturbations.

DISCUSSION

This study offers one of the most granular and semantically rich mappings to date of how Circular Economy (CE) and Big Data Analytics (BDA) research have evolved over the past decade. By combining transformer-based embeddings (SciBERT), clustering (UMAP–HDBSCAN), probabilistic topic modeling (LDA), and embedding-based validation (Top2Vec), the analysis provides a comprehensive view of both thematic structure and temporal development. The findings reveal several important patterns that both confirm and extend existing knowledge in CE digitalization research.

Technical / Industry 4.0 & Analytics Dominance

A key finding is the predominance of Operational Practices and Technical Enablers, which together account for more than half of the topics extracted. This distribution is in line with research on CE that focuses on material recovery, recycling, and manufacturing innovation [5], as well as the enabling role of digital infrastructures [1,24]. Previous bibliometric studies and thematic analyses also show strong clusters around waste valorisation [64], Industry 4.0-enabled circular processes, and digital manufacturing analytics and servitization mechanisms [65]. On a macro level, this thematic formation is consistent with recent CE research frameworks, which frame technological enablers, digitalization, and process optimization as critical tools for operationalizing circular strategy, particularly via Industry 4.0, advanced analytics, and data-driven decision support [41,59,60]. This trajectory is evident in the dominance of analytics- and technology-based topics in the CE–BDA corpus, where technologies are often portrayed as instrumental tools for achieving operational circularity.

Waste, Materials & Environmental Process Optimization

The ontology–community mapping further shows that CE–BDA research concentrates heavily on waste management, material recovery, chemical recycling, and resource-efficiency optimization. This is consistent with the industrial ecology foundations of CE, which position operational transformation at the core of circular transition strategies [66].

In parallel, technical enablers are deeply embedded within these operational domains—IoT, blockchain, AI, and hyperspectral sensing appear alongside recycling, sorting, and manufacturing applications—reflecting the shift toward "smart circularity" emphasized in recent CE–Industry 4.0 frameworks [12, 67].

Business Models & Management: Peripheral but Strategically Critical

Although the impact of business-model innovation in scaling circular solutions is recognized, Business Models and management-oriented themes seem to remain on the periphery of the CE–BDA topic structure. While such narratives often promote servitization and data-driven value creation [65], CE–BDA scholarship tends to emphasize technical development and operational optimization rather than organizational redesign and value-capture mechanisms.

Supply Chain & Operations: Narrow Integration Relative to Technical Depth

Supply chain and operational coordination themes are present but comparatively constrained relative to technical and waste-centric domains. This indicates that CE–BDA research often prioritizes localized analytics and process optimization over system-wide integration and multi-actor coordination—an issue that matters because circular value loops are inherently networked and inter-organizational.

Policy, Governance & Social Dimensions: Persistent Gap and Publication-Channel Effects

In contrast to previous foundational reviews that emphasize policy, governance, and social innovation as significant domains of CE [41,60], the ontology mapping suggests that policy/social topics remain marginal in CE when considered exclusively in a big-data context. Governance, regulation, or societal engagement were the only themes linked with 6 of 88 topics. This divergence indicates that although CE policy research is widely practiced across the broader literature, CE–BDA work is still mainly technical and operational. Perhaps publication-channel bias is partly a reason. The moderate Gini coefficient for journal distribution (0.43) suggests non-trivial convergence of CE–BDA research in technically oriented sustainability publications, which can amplify operational-technical themes while neglecting governance-oriented ideas. Policy and practice implications: Insufficient focus on governance suggests that technological maturity may be outpacing institutional readiness. This strengthens the case for data governance frameworks that address interoperability, traceability standards, auditability, access control, and accountability—especially in multi-actor circular ecosystems, where data sharing and verification are key to scaling circular interventions.

Critical Comparison to SOTA: Fragmentation Versus "Unified" Digital CE Narratives

The topic modelling studies in CE usually extract eight to thirty general clusters, which predominantly focus on waste management, business models, sustainable development, and life-cycle perspectives [39,40]. It is also becoming common for reviews to focus on the

effects of digital technologies—Industry 4.0, IoT, AI, blockchain, and analytics—in CE transitions [68, 69]. Yet, despite the emphasis on digitalization, the majority of empirical work remains at a rather coarse-grained level and does not measure fragmentation based on micro-topic counts, semantic distances, community structure, or ontology alignment [31, 37, 42]. Where the recent CE digitalization syntheses often represent convergent paths towards integrated "smart circularity" frameworks [68,69], the current studies offer empirical evidence of the specialization and fragmentation in CE–BDA. The presence of 88 semantically coherent micro-topics, grouped into distinct communities, indicates parallel technical routes and limited integration outside governance and socio-institutional domains. This indicates that "digital circularity" is evolving through specialized, localized application niches rather than through coordinated socio-technical consolidation, with consequences including the potential for scalability, comparability of evidence, and harmonized policies.

Model Robustness and Interpretability: Why the Topic Structure is Trustworthy

The minimal overlap of keywords across the majority of topic pairs provides empirical evidence of the strength of the BERTopic–SciBERT architecture. Overlap is often associated with semantic leakage and topic instability, especially in high-dimensional, interdisciplinary corpora [55,70,71]. By contrast, sparse similarity structures are most often considered indicators for highly divergent and semantically consistent constructs. The largely sparse overlap heatmap suggests that this configuration helps avoid thematic redundancy, which is a common problem with embedding-only clustering methods applied to scientific corpora [31, 53]. The inclusion of domain-adapted SciBERT embeddings with class-based TF–IDF weighting probably alleviates the impact of contextual embeddings that highlight generic similarity [30, 72], thereby increasing the topic's distinctiveness and interpretability. Simultaneously, pockets of moderate overlap in the local context reflect meaningful interdisciplinary connections that are key in CE–BDA research, encompassing the intersection of technical enablers, environmental processes, and organizational practices [59, 60]. Similar controlled-overlap patterns have been observed in analyses of sustainability and Industry 4.0 topics and are seen as meaningful cross-domain integration, not simply modeling artifacts [8, 73]. These results, taken together with the embedding sensitivity and ablation experiments, reflect a balanced trade-off between thematic specificity and cross-domain compatibility. Divergences in UMAP dimensionality were observed to yield different cluster sizes in the geometry and outlier proportions, but no significant shifts in coherence at the 0.05 level, providing evidence for structural stability. A general-purpose embedding model, in place of SciBERT, also decreased semantic coherence, demonstrating the importance of a domain-adapted model for sustainability corpora. Top2Vec also provides complementary validation—it yielded only two dominant topics, indicating the limitations of embedding-only density clustering in interdisciplinary domains. Top2Vec over-aggregates semantically diverse streams into overly coarse structures by using joint document–word embeddings without lexical reweighting or hierarchical control [53, 74]. This result provides support for hybrid

approaches that integrate domain-adapted embeddings, density clustering, and cTF–IDF-style lexical reweighting to capture semantic complexity in CE–BDA work.

Temporal Dynamics and Structural Shifts

In time, the temporal trends of thematic prevalence from 2016 to 2025 are statistically significant (χ^2 $p < 0.05$). However, technology enablers (AI, IoT, blockchain, and analytics) increase significantly after 2018, when Industry 4.0 enters the field and sustainability policy acceleration is on the agenda. The alignment between past studies and reviews of digital transformation suggests that 2018–2020 was a key inflection point in the adoption of CE technology [1, 12]. A change-point detection reveals structural breaks around 2019–2021, suggesting a transition from foundational capability development to application-driven CE analytics. This aligns with data indicating that sustainability analytics has shifted from exploratory digitalization to operational optimization [75-81]. This interpretation is further supported by the growth-rate patterns, with high-growth themes largely in applied analytics and conceptual and governance themes remaining comparatively stagnant.

Synthesis: Mature in Technology, Immature in Governance (SDG Lens)

As a whole, the findings indicate that CE–BDA scholarship is transitioning toward operational maturity, albeit governance, social aspects, and business-model innovation remain underdeveloped. This is consistent with concerns that CE transitions risk becoming overly technocentric without accompanying institutional and behavioral transformation [41, 66]. From a sustainability perspective, this imbalance shows that CE–BDA research strongly supports SDG 9 (Industry, Innovation and Infrastructure), but insufficient attention to governance and social adoption may hinder efforts toward SDG 12 (Responsible Consumption and Production) and SDG 17 (Partnerships for the Goals). Accordingly, our findings not only support but quantitatively demonstrate a structural imbalance in CE–BD scholarship. In summary, we conclude from the discussion that CE–BDA scholarship has reached a relatively mature phase in both technical and operational fields, while institutional, governance, and socio-behavioral aspects remain comparatively underdeveloped. This disparity, together with the empirically evidenced fragmentation into specialized micro-domains and the validated robustness of the extracted topic structure, compels the summary of contributions, limitations, and future research agenda of the concluding section, on closing governance and business-model gaps while advancing cross-domain integration. The integration of evidence by hypothesis testing, robustness checks, and structural analyses suggests that CE–BDA scholarship is developing significantly in technical and operational aspects. At the same time, governance and socio-institutional dimensions remain systematically underdeveloped, which informs the targeted recommendations and research roadmap offered in the Conclusions.

Advances over the State of the Art and Remaining Research Gaps in CE–BDA

While initial expectations (H4) suggested strong alignment between ontology classes and detected topic communities, statistical testing revealed only moderate agreement (Cramér's $V = 0.318$; $p = 0.91$). This weak alignment implies that existing CE ontologies—

though useful for structuring conceptual domains may not fully capture the emergent structure of CE–BDA research as detected via data-driven methods. In particular, interdisciplinary and hybrid topics may span multiple ontology domains, resulting in fragmented or ambiguous mappings. This suggests a pressing need to evolve CE ontologies to reflect the increasing convergence of digital enablers, operational analytics, and sustainability practices. Rather than undermining the framework, this finding provides a realistic lens on the limitations of current conceptual models when applied to fast-evolving digital CE research.

Overall, this study advances the methodological and conceptual state of the art in Circular Economy–Big Data research by combining transformer-based topic modeling, ontology-driven interpretation, and network-based structural validation. At the same time, the identified gaps point toward future research directions, including the development of CE-specific semantic ontologies, standardized evaluation benchmarks for neural topic models, and deeper integration of policy and governance perspectives into data-driven circular-economy scholarship.

Table 12. Advances over State-of-the-Art and Remaining Gaps in CE–BD Topic Modeling

Dimension	Advances over Prior SOTA	Remaining Gaps / Open Challenges
Topic modeling methodology	First large-scale application of transformer-based topic modeling (BERTopic + SciBERT) to CE–BDA literature, enabling fine-grained and semantically coherent topic extraction beyond LDA-based reviews	Lack of standardized benchmarking protocols for comparing neural topic models across heterogeneous CE corpora
Semantic resolution	Identification of 88 interpretable micro-topics, capturing operational, technological, and environmental CE dimensions with high lexical diversity	Topic granularity remains sensitive to corpus composition and abstraction level; transferability to full-text corpora remains underexplored
Temporal analysis	Empirical evidence of a post-2019 structural shift toward data-driven and operational CE research, supported by χ^2 testing and change-point diagnostics	Limited understanding of causal drivers behind thematic shifts (e.g., policy shocks, funding priorities, global crises)
Ontology integration	Ontology-based topic categorization aligned with Ellen MacArthur Foundation constructs, bridging computational modeling and CE theory	Only moderate alignment found (Cramér's V = 0.318), suggesting that current CE ontologies may not fully capture data-driven topic clusters, especially those at the intersection of digital and operational domains.

Community structure	Partial alignment between topic communities and ontology classes observed qualitatively, but chi-square test indicates no statistically significant association ($p = 0.91$)	Community detection remains dependent on resolution parameters and similarity thresholds
Policy and governance insights	Quantitative confirmation of systematic underrepresentation of policy/social CE themes, highlighting maturity gaps in the field	Absence of integrated socio-technical frameworks linking policy, business models, and operational CE analytics
Methodological robustness	Sensitivity analysis and embedding ablation confirm robustness of findings under reasonable modeling perturbations	Lack of consensus on best practices for robustness validation in neural topic modeling studies

SUMMARY AND CONCLUSIONS

This study provides one of the most fine-grained and methodologically rigorous semantic analyses to date of the intersection between Circular Economy (CE) and Big Data Analytics (BDA). By leveraging domain-adapted transformer embeddings (SciBERT) within BERTopic (UMAP–HDBSCAN and cTF–IDF), complemented by probabilistic modeling (LDA) and embedding-based validation (Top2Vec), the research moves beyond traditional co-word and LDA-centric reviews to reveal the micro-structure of CE digitalization research. The resulting 88 semantically coherent topics, organized into two major communities and mapped to five circular economy ontology classes, provide a high-resolution representation of how data-driven technologies are shaping contemporary CE scholarship. However, statistical testing did not confirm strong structural alignment between community clusters and ontology categories, indicating that data-driven topic groupings only partially reflect conceptual taxonomies

Findings show that Operational Practices and Technical Enablers dominate most CE–BDA research, which confirms the central role that digital technologies like IoT, AI, blockchain, hyperspectral sensing, and analytics play in enabling circular strategies. In contrast, Policy/Social and Business Model themes remain comparatively marginal, indicating a persistent structural imbalance between technological innovation and institutional, behavioral, and governance dimensions. Temporal evidence further indicates a statistically significant shift around 2019–2021, consistent with the acceleration of Industry 4.0 diffusion and the growing emphasis on data-driven sustainability practices.

One valuable aspect of this study is the provision of computational evidence that CE digitalization research is fragmenting into increasingly specialized micro-domains. Although previous CE literature typically identifies 8–30 broad themes, we identify 88 micro-topics capturing fine-grained sub-domains (e.g., hyperspectral waste sorting, blockchain-enabled energy optimization, photovoltaic recycling analytics). Such granularity builds on previous CE digitalization work that acknowledges increasing

specialization but does not quantify the number of emerging micro-topics, their semantic separation, community structure, or ontology alignment [68,69]. By contrast, the present work quantifies fragmentation through the integrated use of SciBERT-based embeddings, Louvain community detection, ontology mapping (Ellen MacArthur Foundation categories), temporal evolution analysis, and statistically significant variation testing (χ^2). The study thus represents a move beyond largely narrative observation towards reproducible, data-driven characterization of CE–BDA thematic specialization.

To practitioners as well as policymakers, the findings reveal where digital CE innovation is currently concentrated, particularly in manufacturing, recycling, and energy, and highlight critical blind spots in governance, user engagement, and social impact assessment. Bridging these gaps is crucial to guarantee that digital changes undergird a holistic and inclusive circular transition rather than reinforcing a narrowly techno-centric pathway.

The dataset, modeling workflow, and analytical framework provide a replicable basis for computational CE studies that researchers can perform in future work. Future work could contribute to this study by (i) applying full-text corpora and domain-specific language models to extend socio-institutional coverage, (ii) extending temporal windows and comparing regional trajectories, and (iii) integrating semantic structures with citation networks, geographic trends, policy uptake, or quantifiable sustainability outcomes.

Our findings highlight not only strengths but also limitations in existing CE ontologies, which may lag behind the thematic complexity and hybridization visible in data-driven topic clusters. This calls for further development of ontological frameworks that better integrate digital and interdisciplinary constructs in CE research.

In summary, this study establishes a benchmark for semantic, temporal, and structural analysis of CE–BDA research. By providing high-resolution evidence of thematic fragmentation and identifying areas of persistent underrepresentation—especially in governance and business-model innovation—it contributes, both theoretically and practically, to advancing the next generation of circular-economy digitalization research.

LIMITATIONS AND FUTURE WORK

Methodological Limitations

First, while SciBERT provides strong representations for scientific language, it was not fine-tuned on a CE-specific corpus. As a result, some highly specialized circular-economy terminology or context-dependent meanings may be only partially captured. Second, although the BERTopic pipeline (UMAP–HDBSCAN with cTF–IDF) is well-suited to large scholarly corpora, the resulting topic structure can remain sensitive to density-based clustering hyperparameters; alternative clustering formulations (e.g., hierarchical or probabilistic variants) could yield additional sub-themes or different boundary definitions. Third, the Top2Vec baseline converged to a very small number of coarse topics in this interdisciplinary setting, reinforcing evidence that embedding-only density clustering may

over-aggregate heterogeneous research streams when lexical reweighting and topic representation controls are not incorporated.

Data and Coverage Limitations

The dataset is limited to English-language publications indexed in Scopus and Web of Science. Consequently, regional or non-indexed research outputs may be underrepresented, and grey literature (e.g., policy reports, standards, industrial guidelines) is excluded despite its relevance to governance and implementation. In addition, the primary reliance on abstracts rather than full texts may reduce sensitivity to methodological detail, contextual nuance, and application-specific constraints that are often contained in body sections (e.g., data pipelines, evaluation protocols, deployment limitations).

Conceptual and Interpretive Limitations

The ontology mapping, while grounded in established CE theory, necessarily simplifies emerging and hybrid CE–digitalization domains. Some topics plausibly span multiple ontology categories (e.g., AI-enabled governance mechanisms or data-driven servitization), and a single-label assignment does not fully capture this cross-category structure.

Directions for Future Work and Mitigation

Future research should (i) evaluate domain-tuned language models (or lightweight CE-adaptation of SciBERT), (ii) incorporate full-text and multimodal evidence where feasible, and (iii) integrate semantic topic structures with citation, co-authorship, and geographic networks to characterize intellectual influence and diffusion. A targeted mitigation for the abstract-based limitation is to sample and analyze a subset of full-text articles (e.g., 50 items) and assess whether the same dominant topics and macro-cluster structure are recovered, thereby validating the stability of abstract-derived themes. Finally, governance, policy, and business-model dimensions—currently underrepresented in the CE–BDA corpus—remain priority areas where future big-data applications could materially enrich CE scholarship.

AUTHORS CONTRIBUTIONS

Conceptualization, EM, IF, and EP; Methodology, IF and EP; Software and Computational Modeling, EP; Validation, IF, EM, and EP; Formal Analysis, IF and EP; Investigation, EM and IF; Resources, IF; Data Curation, IF and EP; Writing – Original Draft Preparation, EP; Writing – Review & Editing, IF and EM; Visualization, EP; Supervision, EM;

CONFLICT OF INTERESTS

The authors should confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Neri, A., Cagno, E., Susur, E., Urueña, A., Nuur, C., Kumar, V., et al. The relationship between digital technologies and the circular economy: a systematic literature review and a research agenda. *R&D Management* **2025**, 55, 617–713.
2. Edwin Cheng, T.C., Kamble, S.S., Belhadi, A., Ndubisi, N.O, Lai, K., Kharat, M.G. Linkages between big data analytics, circular economy, sustainable supply chain flexibility, and sustainable performance in manufacturing firms. *Int J Prod Res* **2022**, 60, 6908–22.
3. Dayal U, Gupta M, Ghosh D, Wadhawan D, Morrow A, Horiguchi S, et al. Enabling Product Circularity Through Big Data Analytics and Digitalization. *2022 IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS), IEEE*; **2022**, p. 1–6.
4. Mamudu, U.U., Obasi, C.D., Awuye, S.K., Danso, H., Ayodele, P., Akinyemi, P. Circular economy in the manufacturing sector: Digital transformation and sustainable practices. *International Journal of Science and Research Archive* **2024**, 12, 129–41.
5. Kristoffersen, E., Blomsma, F., Mikalef, P., Li, J. The smart circular economy: A digital-enabled circular strategies framework for manufacturing companies. *J Bus Res* **2020**, 120, 241–61.
6. Sasso, R.A., Filho, M.G., Ganga, G.M.D. Synergizing lean management and circular economy: Pathways to sustainable manufacturing. *Corp Soc Responsib Environ Manag* **2025**, 32, 543–62.
7. Siratan, E.D., Anatasia, V., Arifandi, A., Setiadi, H., Kriswanto, D. Optimizing Sustainable Supply Chain Management with a Circular Economy Approach in the Manufacturing Industry. *International Journal for Science Review* **2025**, 2. 183-193
8. Awan, U., Sroufe, R., Shahbaz, M. Industry 4.0 and the circular economy: A literature review and recommendations for future research. *Bus Strategy Environ* **2021**, 30, 2038–60.
9. Beltagy, I., Lo, K., Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the *9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics; **2019**, p. 3613–8.
10. Frank, A.G., Dalenogare, L.S., Ayala, N.F. Industry 4.0 technologies: Implementation patterns in manufacturing companies. *Int J Prod Econ* **2019**, 210, 15–26.
11. Plekhanov, D., Franke, H., Netland, T.H. Digital transformation: A review and research agenda. *Eur Management Journal* **2023**, 41, 821–44.
12. Toth-Peter, A., de Oliveira, R.T., Mathews, S., Barner, L., Figueira, S. Industry 4.0 AS an enabler in transitioning to circular business models: A systematic literature review. *J Clean Prod* **2023**, 393, 136284.
13. Dolci, V., Bigliardi, B., Petroni, A., Pini, B., Filippelli, S., Tagliente, L. Integrating Industry 4.0 and Circular Economy: A Conceptual Framework for Sustainable Manufacturing. *Procedia Comput. Sci.* **2024**, *232*, 1711–1720.
14. Sahu, A., Agrawal, S., Kumar, G. Integrating Industry 4.0 and circular economy: a review. *J Enterp Inf Manag* **2021**, 35, 885–917.
15. Reis, W.F., Barreto, C.G., Capelari, M.G.M. Circular Economy and Solid Waste Management: Connections from a Bibliometric Analysis. *Sustainability* **2023**, 15(22), 15715.
16. Saha K, Farhanj Z, Kumar V. A systematic review of circular economy literature in healthcare: Transitioning from a 'post-waste' approach to sustainability. *J Clean Prod* **2025**, 505, 145427.

17. Shakhnoza Nuralievna, K., Park, Y.W. Circular Economy in Automobile Industry: Literature Analysis. *KINFORMS* **2023**, 18, 83–113.
18. Hazen, B.T., Russo, I., Confente, I., Pellathy, D. Supply chain management for circular economy: conceptual framework and research agenda. *The International Journal of Logistics Management* **2021**, 32, 510–37.
19. Montag, L. Circular Economy and Supply Chains: Definitions, Conceptualizations, and Research Agenda of the Circular Supply Chain Framework. *Circular Economy and Sustainability* **2023**, 3, 35–75.
20. Zhang, A., Duong, L., Seuring, S., Hartley, J.L. Circular supply chain management: a bibliometric analysis-based literature review. *The International Journal of Logistics Management* **2023**, 34, 847–72.
21. Guillén-Pacho, I., Badenes-Olmedo, C., Corcho, O. Dynamic topic modelling for exploring the scientific literature on coronavirus: an unsupervised labelling technique. *Int J Data Sci Anal* **2025**, 20, 2551–81.
22. Garcia-Muiña, F.E., González-Sánchez, R., Ferrari, A.M., Settembre-Blundo, D. The Paradigms of Industry 4.0 and Circular Economy as Enabling Drivers for the Competitiveness of Businesses and Territories: The Case of an Italian Ceramic Tiles Manufacturing Company. *Soc. Sci.* **2018**, *7*(12), 255.
23. Mousa Mousa, M., Abdulrahman Al Moosa, H., Naim Ayyash, I., Omeish, F., Zaiem, I., Alzahrani, T., Hammami, S.M., Zamil, A.M. Big Data Analytics as a Driver for Sustainable Performance: The Role of Green Supply Chain Management in Advancing Circular Economy in Saudi Arabian Pharmaceutical Companies. *Sustainability* **2025**, *17*, 6319.
24. Tavera Romero, C.A., Castro, D.F., Ortiz, J.H., Khalaf, O.I., Vargas, M.A. Synergy between Circular Economy and Industry 4.0: A Literature Review. *Sustainability* **2021**, 13, 4331
25. Ingemarsdotter, E., Jamsin, E., Kortuem, G., Balkenende, R. Circular Strategies Enabled by the Internet of Things—A Framework and Analysis of Current Practice. *Sustainability* **2019**, 11, 5689.
26. Rajput, S., Singh, S.P. Connecting circular economy and industry 4.0. *Int J Inf Manage* **2019**, 49, 98–113.
27. Gan, L. *et al.* Experimental Comparison of Three Topic Modeling Methods with LDA, Top2Vec and BERTopic. In: Lu, H., Cai, J. (eds) Artificial Intelligence and Robotics. *ISAIR 2023. Communications in Computer and Information Science*, vol 1998. Springer, Singapore. **2024**, p. 376–91.
28. Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., Hemphill, L. A Bibliometric Review of Large Language Models Research from 2017 to 2023. *ACM Trans Intell Syst Technol* **2024**, *15*(5), 1-25.
29. Zupic, I., Čater, T. Bibliometric methods in management and organization. *Organ Res Methods* **2015**, *18*, 429–72.
30. Grootendorst, M. BBERTopic: Neural topic modeling with a class-based TF-IDF procedure. **2022**.
31. Egger, R., Yu, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology* **2022**, *7*, 1-18.
32. Ogunleye, B., Lancho Barrantes, B.S., Zakariyyah, K.I. Topic modelling through the bibliometrics lens and its technique. *Artif Intell Rev* **2025**, *58*, 74.
33. Voskergian, D., Jayousi, R., Yousef, M. Topic selection for text classification using ensemble topic modeling with grouping, scoring, and modeling approach. *Sci Rep* **2024**, *14*, 23516.

34. Compton, T. Beyond the Black Box: Integrating Lexical and Semantic Methods in Quantitative Discourse Analysis with BERTopic **2025**.
35. Sharifian-Attar V, De S, Jabbari S, Li J, Moss H, Johnson J. Analysing Longitudinal Social Science Questionnaires: Topic modelling with BERT-based Embeddings. 2022 *IEEE International Conference on Big Data (Big Data), IEEE*; **2022**, p. 5558–67.
36. Vaccargiu M, Tonelli R. Blockchain Projects in Environmental Sector: Theoretical and Practical Analysis. *Earth* **2024**, 5, 354–70.
37. Raman, R., Das, P., Aggarwal, R., Buch, R., Palanisamy, B., Basant, T., et al. Circular Economy Transitions in Textile, Apparel, and Fashion: AI-Based Topic Modeling and Sustainable Development Goals Mapping. *Sustainability* **2025**, 17(12), 5342.
38. Albrekht, V., Mukhamediev, R.I., Popova, Y., Muhamedijeva, E., Botaibekov, A. Top2Vec Topic Modeling to Analyze the Dynamics of Publication Activity Related to Environmental Monitoring Using Unmanned Aerial Vehicles. *Publications* **2025**, 13, 15.
39. Mahanty, S., Boons, F., Handl, J., Batista-Navarro, R. (2019). Studying the Evolution of the '*Circular Economy*' Concept Using Topic Modelling. In: Yin, H., Camacho, D., Tino, P., Tallón-Ballesteros, A., Menezes, R., Allmendinger, R. (eds) *Intelligent Data Engineering and Automated Learning – IDEAL 2019. IDEAL 2019. Lecture Notes in Computer Science*, vol 11872. Springer. **2019**, p. 259–70.
40. Dragomir, V.D., Dumitru, M. The state of the research on circular economy in the European Union: A bibliometric review. *Cleaner Waste Systems* **2024**, *7*, 100127.
41. Kirchherr, J., Reike, D., Hekkert, M. Conceptualizing the circular economy: An analysis of 114 definitions. *Resour Conserv Recycl* **2017**, *127*, 221–32.
42. Rejeb, A., Rejeb, K., Keogh, J.G., Süle, E. When Industry 5.0 Meets the Circular Economy: A Systematic Literature Review. *Circular Economy and Sustainability* **2025**, *5*, 2621–52.
43. Giordano, V., Castagnoli, A., Pecorini, I., Chiarello, F. Identifying technologies in circular economy paradigm through text mining on scientific literature. *PLoS One* **2024**, *19*, e0312709-.
44. Surian, D., Nguyen, D.Q., Kennedy, G., Johnson, M., Coiera, E., Dunn, A.G. Characterizing Twitter Discussions About HPV Vaccines Using Topic Modeling and Community Detection. *J Med Internet Res* **2016**, 18, e232.
45. Pandis, N. The chi-square test. *American Journal of Orthodontics and Dentofacial Orthopedics* **2016**, *150*, 898–9.
46. McInnes, L., Healy, J., Saul, N., Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw* **2018**, *3*, 861.
47. Campello, R.J.G.B, Moulavi, D., Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G, editors. *Advances in Knowledge Discovery and Data Mining*, Berlin, Heidelberg: Springer Berlin Heidelberg; **2013**, p. 160–72.
48. Asyaky, M.S., Mandala, R. Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP. 2021 *8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, IEEE; **2021**, p. 1–6.
49. Eklund, A., Forsman, M. Topic Modeling by Clustering Language Model Embeddings: Human Validation on an Industry Dataset. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Stroudsburg, PA, USA: Association for Computational Linguistics; 2022, p. 635–43.

50. Navarro, E.G., Homayouni H. Topic Modeling in Cardiovascular Research Publications, **2023**.
51. Bianchi, F., Terragni, S., Hovy, D. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics; **2021**, p. 759–66.
52. Blei, D.M., Ng, A.Y., Jordan, M.I. Latent dirichlet allocation. *J Mach Learn Res* **2003**, *3*, 993–1022.
53. Angelov, D. Top2Vec: Distributed Representations of Topics **2020**.
54. Aria, M., Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *J Informetr* **2017**, 11, 959–75.
55. Blei, D.M. Probabilistic topic models. *Commun ACM* **2012**, *55*, 77–84.
56. Callon, M., Courtial, J.P., Laville, F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics* **1991**, *22*, 155–205.
57. Hannigan, T.R., Haans, R.F.J, Vakili, K., Tchalian, H., Glaser, V.L., Wang, M.S, et al. Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals* **2019**, *13*, 586–632.
58. van Eck, N.J., Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523–38.
59. Bocken, N.M.P., de Pauw, I., Bakker, C., van der Grinten, B. Product design and business model strategies for a circular economy. *Journal of Industrial and Production Engineering* **2016**, 33, 308–20.
60. Geissdoerfer, M., Savaget, P., Bocken, N.M.P., Hultink, E.J. The Circular Economy – A new sustainability paradigm? *J Clean Prod* **2017**, 143, 757–68.
61. Ellen MacArthur Foundation. Towards A Circular Economy: Business Rationale for an Accelerated Transition. **2015**.
62. Killick, R., Fearnhead, P., Eckley, I.A. Optimal Detection of Changepoints with a Linear Computation Cost. *J Am Stat Assoc* **2012**, *107*, 1590–8.
63. Kirchherr, J., Reike, D., Hekkert, M. Conceptualizing the circular economy: An analysis of 114 definitions. *Resour Conserv Recycl* **2017**, *127*, 221–32.
64. Govindan, K., Hasanagic, M. A systematic review on drivers, barriers, and practices towards circular economy: a supply chain perspective. *Int J Prod Res* **2018**, *56*, 278–311.
65. Bressanelli, G., Adrodegari, F., Perona, M., Saccani, N. Exploring How Usage-Focused Business Models Enable Circular Economy through Digital Technologies. *Sustainability* **2018**, *10*, 639
66. Korhonen, J., Honkasalo, A., Seppälä, J. Circular Economy: The Concept and its Limitations. *Ecolog Eco* **2018**, *143*, 37–46.
67. Tseng, M., Ha, H.M., Tran, T.P.T., Bui, T., Chen, C., Lin, C. Building a data-driven circular supply chain hierarchical structure: Resource recovery implementation drives circular business strategy. *Bus Strategy Environ* **2022**, *31*, 2082–106.
68. Khan, A.A., Abonyi, J. Information sharing in supply chains – Interoperability in an era of circular economy. *Cleaner Logistics and Supply Chain* **2022**, 5, 100074.
69. Trevisan, A.H., Zacharias, I.S., Liu, Q., Yang, M., Mascarenhas, J. Circular Economy and Digital Technologies: A Review of the Current Research Streams. *Proceedings of the Design Society* **2021**, 1, 621–30.

70. Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A. Optimizing Semantic Coherence in Topic Models. In: Barzilay R, Johnson M, editors. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK.: Association for Computational Linguistics; **2011**, p. 262–72.
71. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D. Exploring Topic Coherence over Many Models and Many Topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea: Association for Computational Linguistics; **2012**, p. 952–61.
72. Sia, S., Dalmia, A., Mielke, S.J. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics; **2020**, p. 1728–36.
73. Matarneh, S., Piprani, A.Z., Ellahi, R.M., Nguyen, D.N., Mai Le, T., Nazir, S. Industry 4.0 technologies and circular economy synergies: Enhancing corporate sustainability through sustainable supply chain integration and flexibility. *Environ Technol Innov* **2024**, *35*,103723.
74. Kim, K., Kogler, D.F., Maliphol, S. Identifying interdisciplinary emergence in the science of science: combination of network analysis and BERTopic. *Humanit Soc Sci Commun* **2024**, *11*, 603.
75. Shehu, M., Stringa, A., & Gjika Dhamo, E. A Latent Dirichlet Allocation Framework to Analyse and Forecast Employability Skills. *Int. Journ. of Innov. Techn. and Interdisc. Sci.*, **2025**, *8*(3), 595–623.
76. Nargiz, H., Nushaba, H., & Ayshan, M. Development of Bank Marketing in the Conditions of Digital Transformation in Azerbaijan. *Journal of Integrating Engineering and Applied Sciences*, **2025**, *3*(2), 217–228.
77. Kaprata, G., & Kume, A. The Garden City as an Urban Paradigm for a Sustainable Economic Model: The Case of Albania in the Fourth Post-Communist Decade. *International Journal of Innovative Technology and Interdisciplinary Sciences*, **2025**, *8*(4), 994–1020.
78. Prifti V., Markja, I., Dhoska, K., Pramono, A. *IOP Conf. Ser.: Mater. Sci. Eng.* **2020**, *909*, 012047
79. Gheibi, M., Masoomi, S. R., Magala, M. U., Nalugo, D., & Kassim, A. T. H. (2024). Motivational Electronic Waste Management System in CXI, Technical University of Liberec. *Journal of Transactions in Systems Engineering*, *3*(1), 325–339.
80. Prifti, K., Vrusho, B., Toci, Çlirim, Prendi, L., & Bushi Gjuzi, J. Strategic Human Resource Management and Its Impact on Organizational Performance: Empirical Insights. *International Journal of Innovative Technology and Interdisciplinary Sciences*, **2025**, *8*(3), 550–594.
81. Esposito, M., Tse, T., Soufani, K. Introducing a Circular Economy: New Thinking with New Managerial and Policy Implications. *Calif Manage Rev* **2018**, *60*, 5–19.