future internet

MDPI

Article

Investigating the Key Aspects of a Smart City through Topic Modeling and Thematic Analysis

Anestis Kousis and **Christos Tjortjis** *

Department of Science and Technology, International Hellenic University, 14th km Thessaloniki-N. Moudania National Road, 57001 Thermi, Greece; a.kousis@ihu.edu.gr
* Correspondence: c.tjortjis@ihu.edu.gr; Tel.: +30-2310-807-576

Abstract: In recent years, the emergence of the smart city concept has garnered attention as a promising innovation aimed at addressing the multifactorial challenges arising from the concurrent trends of urban population growth and the climate crisis. In this study, we delve into the multifaceted dimensions of the smart city paradigm to unveil its underlying structure, employing a combination of quantitative and qualitative techniques. To achieve this, we collected textual data from three sources: scientific publication abstracts, news blog posts, and social media entries. For the analysis of this textual data, we introduce an innovative semi-automated methodology that integrates topic modeling and thematic analysis. Our findings highlight the intricate nature of the smart city domain, which necessitates examination from three perspectives: applications, technology, and socio-economic perspective. Through our analysis, we identified ten distinct aspects of the smart city paradigm, encompassing mobility, energy, infrastructure, environment, IoT, data, business, planning and administration, security, and people. When comparing the outcomes across the three diverse datasets, we noted a relative lack of attention within the scientific community towards certain aspects, notably in the realm of business, as well as themes relevant to citizens' everyday lives, such as food, shopping, and green spaces. This work reveals the underlying thematic structure of the smart city concept to help researchers, practitioners, and public administrators participate effectively in smart city transformation initiatives. Furthermore, it introduces a novel data-driven method for conducting thematic analysis on large text datasets.

Keywords: smart city; topic modeling; thematic analysis; BERTopic

check for updates

Citation: Kousis, A.; Tjortjis, C. Investigating the Key Aspects of a Smart City through Topic Modeling and Thematic Analysis. *Future Internet* **2024**, *16*, 3. https://doi.org/10.3390/fi16010003

Academic Editor: Antonio Esposito

Received: 27 November 2023
Revised: 15 December 2023
Accepted: 18 December 2023
Published: 22 December 2023

1. Introduction

Contemporary cities face ongoing difficulties in coping with escalating population density and swift urbanization. According to reports, approximately one-third of this population growth is concentrated within urban areas [1], and, as projected by the United Nations [2], the urban populace is anticipated to encompass 68% of the worldwide population by 2050. Numerous municipal administrations have embarked on ambitious initiatives for smart city development, aiming to tackle an array of urban challenges, such as traffic congestion, energy deficits, and ecological contamination spurred by the rapid advance of urbanization and heightened resource demands [3].

However, changing circumstances demand innovative reactions. This is particularly relevant to the expansion of urbanization and the emergence of information and communication technologies (ICT), which are significantly reshaping the landscape of sustainable urbanism. The utilization of emerging technologies, including communication technologies like the Internet of Things (IoT), in conjunction with artificial intelligence (AI), serves this purpose. The combination of AI and IoT has demonstrated promising outcomes in previous studies [4].

In this context, smart city initiatives have been established globally and have garnered substantial attention from both the research and industrial communities over the last decade [5]. The term “smart city” refers to a novel urban environment designed for optimal performance through the integration of ICT and other forms of physical infrastructure [6]. Townsend defines the smart city as “a place where information technology is combined with infrastructure, architecture, everyday objects, and even our bodies to address social, economic, and environmental problems” [7]. The importance of smart cities lies not only in the technology they employ but also in their innovative approach to addressing urban challenges, improving the quality of life for city residents, and optimizing governmental functions [8].

As interest in smart cities grows, the number of scientific articles being published in this field is rapidly increasing [9], and more valuable information is brought together about how a smart city should be designed, developed, and operated. There is an agreement that this information gathering may help in theory creation by prescribing generalizable models that can be validated by data to describe a complex phenomenon, such as a city [10]. In the context of smart cities, generating analytical outcomes can result in more sophisticated understandings of urban phenomena, thereby aiding the development of evidence-driven urban strategies and fostering innovation. Data science technologies can contribute in various ways to this intention, as elaborated in detail in [9].

In the past, scholars employed qualitative methods to analyze textual, unstructured material for complex phenomena understanding in several fields. Qualitative research entails a thorough examination of texts, leading to a deep human understanding and interpretation [11]. However, these manual techniques are not only time-consuming and labor-intensive, but they also present issues owing to coding bias and a lack of uniformity in objectivity [12]. Because of the high number of scientific publications and their rapid expansion, the amount of research is now immense, and its content is so complicated that traditional ways of interpreting the smart city, such as literature reviews, content analysis, and case studies [13], may not be adequate to conduct quantitative analyses, analyze patterns, or make judgments without the assistance of computers. With so many scientific papers available, computer-assisted data summarization is essential [14].

Braun and Clarke [15] suggest thematic analysis as a foundational method for qualitative analysis. They describe thematic analysis as a systematic approach for identifying, analyzing, and reporting recurring patterns or themes inherent in data, facilitating a comprehensive textual exposition. They argue that thematic analysis can extend beyond this, identifying diverse aspects within the domain under investigation. Gillies et al. [11] propose topic modeling as a suitable automated technique for the initial stages of thematic analysis, given that the “themes” of thematic analysis align conceptually with the “topics” in topic modeling. Topic modeling stands as currently the most widely used text-mining technique [16] to explore the interrelation between the gathered data and documents, validating models, condensing the collection of texts, and steering the investigation into its contents [17]. Topic modeling involves the task of uncovering the inherent thematic structure within a text corpus, often resulting in a presentation of the most frequent terms associated with each identified topic [18]. It eliminates the need for researchers to formulate coding sheets prior to analysis, offering a time-efficient method for conducting exploratory assessments on extensive paper collections [19]. Although topic modeling is particularly well-suited for big data, it lacks the subtlety inherent to human interpretation. However, Gillies et al. [11] argue that both approaches share a common objective: identify underlying themes hidden within the data.

This study aims to investigate the underlying structure of the smart city paradigm by employing topic modeling and thematic analysis. It contributes to our understanding of the potential knowledge landscape, particularly within the field of smart city development, by identifying significant themes as well as subthemes. It also facilitates the establishment of a semi-automated novel methodology for understanding the domain. Additionally, smart city leaders can utilize the derived discoveries of the analysis to effectively oversee the smart city structure and tackle challenges within smart city environments. Moreover, we describe a semi-automated methodology for the thematic analysis of large corpora by utilizing a topic modeling technique and further qualitative analysis.

Related Work

Previous research investigating various aspects of the smart city concept has been conducted thus far. Kim et al. [8] introduced a methodology aimed at analyzing diverse informal civic inquiry data for a city and subsequently devising urban plans that align with the preferences of its inhabitants. Through the analysis of approximately 160,000 civic inquiries aimed at contributing to Seoul's sustainable development, the study was intended to facilitate the creation of a citizen-centric and smart city that caters to the preferences of its populace. By employing the dynamic topic model, the authors aimed to pinpoint civic demands and predict the future needs of the citizens. Park and Lee [20] amassed a dataset of 11,527 papers bearing the title "Smart City(-ies)" from databases that included the Scopus database and Springer database. Employing the latent Dirichlet allocation (LDA) topic modeling technique, they proceeded to scrutinize the research landscape, thematic content, and temporal trends, drawing from abstracts and publication year information. Nicolas et al. [3] aimed to enhance comprehension and offer practical directions for future advancements in smart cities. Textual content was gathered from the announcements on the websites of four smart cities. They employed the LDA algorithm to extract strategic topics from the corpus. Following a similarity analysis among the obtained topics and a performance evaluation of the smart cities to compare top-down communication patterns, they identified six pillars of smart city development: smart economy, smart people, smart governance, smart mobility, smart environment, and smart living. Sharma et al. [1] employed an LDA-based analysis on a dataset encompassing 8320 articles spanning the period from 2010 to 2022. They concluded that researchers have prominently concentrated on three key domains: security, connectivity, and decentralization in the context of smart cities. Wang et al. [21] examined the state of research about smart city technologies by conducting a topic modeling analysis on a dataset of 10,000 papers listed in the Web of Science database, spanning from 2009 to 2020. The outcome of the analysis unveiled a categorization of five aspects: policy research on the status quo of smart cities, data analysis and application, infrastructure construction, urban governance, and network security.

Lee [22] applied an LDA topic modeling algorithm to detect emerging topics in smart cities. Subsequently, a temporal analysis of the topic networks was conducted to gain deeper insights into the evolutionary curve of smart city research. Esposito et al. [23] similarly employed LDA on projects proposed by municipal policymakers in Wallonia, Belgium. Their study aimed to investigate the polysemic nature of the smart city concept and underscore the diversity of the prospects presented by smart city policies. Zheng et al. [24] contemplated an instance of utilizing the LDA topic modeling technique on 137 smart city proposals submitted to the Government of Canada's Smart Cities Challenge (SCC) that was operational from 2017 to 2019. Similarly, proposals submitted for the "Inclusive Smart City" project were investigated in [5] through the application of topic modeling techniques.

Alswedani et al. [25] examined the application of AI in urban governance, exploring how AI could aid governments in learning about urban governance parameters to enhance the development of more effective governance strategies. By employing LDA-based topic modeling on Twitter data, they identified ten urban governance parameters. Twitter was also the primary data source for Vargas-Calderón and Camargo's [26] investigation. In their study, they analyzed a sizable dataset comprising 2,634,176 tweets originating from Bogotá's citizens over a period of six months. With the growing number of well-informed citizens actively engaged in urban development, Sinha et al. [27] introduced a system called CUrb. This system is designed to autonomously detect, analyze, and monitor diverse urban-related concerns expressed on social media.

Therefore, the existing literature on revealing the aspects of a smart city through extended textual data has primarily employed traditional techniques such as the LDA algorithm. However, these traditional methods exhibit limitations, particularly when compared to newer approaches like BERTopic. For instance, the LDA algorithm constrains researchers in terms of the number of topics, which must be predetermined, thus limiting the ability to reveal nuanced subtopics. Furthermore, the aforementioned studies typically gathered data from sources such as scientific papers, websites, reports, or social media posts. Notably, none of these studies attempted to compare different sources to highlight the distinctions between top-down and bottom-up approaches. Finally, there is a noticeable gap in the literature concerning further qualitative analysis of the topic modeling results to create a comprehensive thematic map for the smart city.

The remainder of the paper is organized as follows: The next section describes the methodology we adopted for this study. Section 3 addresses the results of the analysis. In Section 4, we discuss our findings, whilst the conclusions are presented in Section 5.

2. Materials and Methods

2.1. Research Design

This study aims to propose a framework for identifying the key aspects of a smart city as they emerge from the thematic analysis of three distinct sources: (a) scientific publications, (b) a business news blog, and (c) a social media platform. The three specific sources were chosen to collect information from academia, businesses, and citizens, respectively. The goal was to establish a foundation for smart city design, development, and management. Specifically, the study begins by investigating emerging topics through an analysis of 13,473 abstracts of pertinent scientific articles, 3224 blog posts, and 2370 social media entries by leveraging a state-of-the-art machine-learning (ML) technique for topic modeling. Subsequently, this paper identifies and compares the topics extracted from the three diverse text corpora, conducting a thorough study of the most representative documents to facilitate a comprehensive interpretation of the investigated domain.

This present article is guided by the following research questions:

RQ1: What are the key aspects of the smart city concept?

RQ2: Which themes within the smart city domain require increased emphasis and attention from researchers?

To achieve our goals, we propose a methodology for investigating the key aspects of a smart city design by employing thematic analysis using both quantitative and qualitative techniques. Many researchers commonly acknowledge that each quantitative and qualitative research technique possesses its own set of strengths and limitations. As a result, it is often recommended to integrate both approaches to ensure that their outcomes harmonize and enhance each other [28].

A typical workflow for thematic analysis is proposed by Braun and Clarke [15], including six phases, as illustrated in Figure 1. They presented a comprehensive guide that entails a recursive process, allowing for movement back and forth as required across the phases. Thematic analysis embarks with a phase of familiarization with the data, during which researchers begin by comprehensively reading the data to develop a holistic understanding of its content before delving into detailed analysis. This is followed by a coding process involving an in-depth examination of the data. Significant excerpts from the data are selected, and "codes" are applied to them. These codes are brief expressions or single words that encapsulate and categorize the text's subject matter. The codes should possess a level of generality that allows them to be applicable across various sections of the text, thus uniting disparate passages discussing the same topic. After completing the initial close coding, the researcher revisits the codes to identify overarching themes by amalgamating codes and exploring their interconnections. Once a series of themes have been extracted, they are assessed and refined by revisiting the original data and comparing them to ensure the alignment of the themes with the data [11]. However, text mining can offer an initial examination of the text, showcasing primary subjects and trends, before engaging in a qualitative assessment of the documents [29].

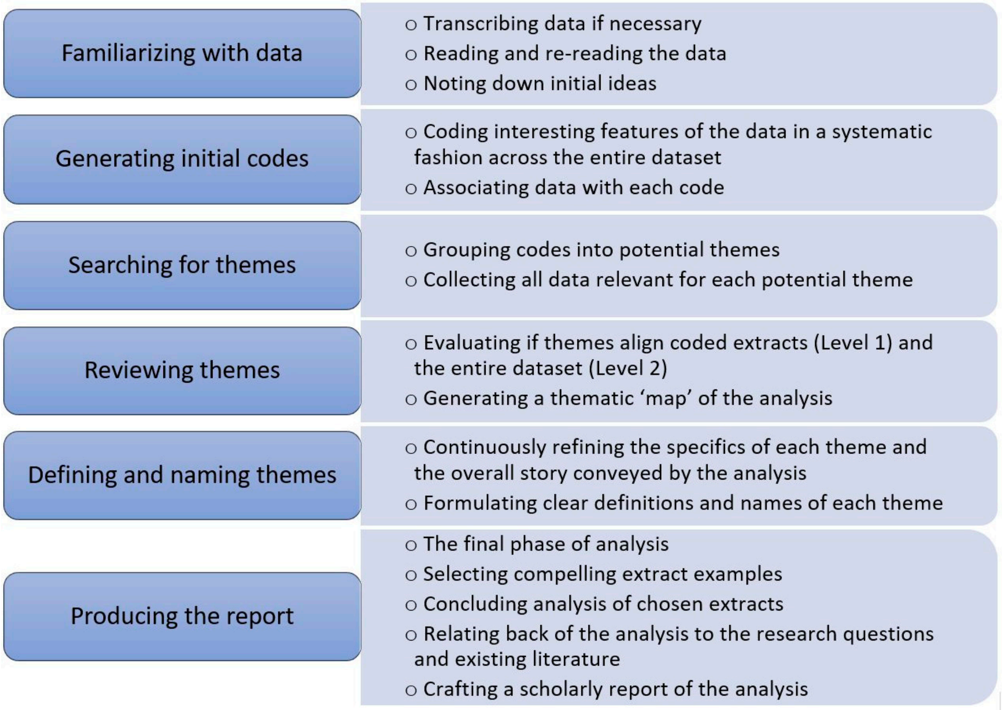
Familiarizing with data
○ Transcribing data if necessary
○ Reading and re-reading the data
○ Noting down initial ideas
Generating initial codes
○ Coding interesting features of the data in a systematic fashion across the entire dataset
○ Associating data with each code
Searching for themes
○ Grouping codes into potential themes
○ Collecting all data relevant for each potential theme
Reviewing themes
○ Evaluating if themes align coded extracts (Level 1) and the entire dataset (Level 2)
○ Generating a thematic 'map' of the analysis
Defining and naming themes
○ Continuously refining the specifics of each theme and the overall story conveyed by the analysis
○ Formulating clear definitions and names of each theme
Producing the report
○ The final phase of analysis
○ Selecting compelling extract examples
○ Concluding analysis of chosen extracts
○ Relating back of the analysis to the research questions and existing literature
○ Crafting a scholarly report of the analysis

Figure 1. Phases of thematic analysis.

The fundamental challenges related to thematic analysis revolve around identifying a theme and determining the number of themes within a given dataset. Topic modeling not only enables the exploration of a massive body of text but also facilitates a more systematic identification of themes. Topics consist of probabilistic word distributions. Grounded in ML and natural language processing (NLP), topic modeling can be employed to depict our documents as probabilistic distributions of topics. Leveraging our awareness of the words and their frequencies in documents, we can utilize this information to construct these topic models. Subsequently, having established our topic model, we can commence portraying all our documents through topic distributions [30]. As a result, there exists a substantial potential for thematic analysis and topic modeling to closely align or even potentially substitute each other, particularly when the objective is to understand the content of a dataset. Both methods are rooted in textual data and yield a comprehensive depiction of the dataset by dissecting it into distinct themes [16]. In our research, we employed a topic modeling approach in conjunction with an in-depth examination of the most relevant documents within each theme. This was done to successfully execute the proposed stages of thematic analysis.

Topic modeling is particularly suitable for the initial three phases of the proposed thematic analysis process, allowing researchers to establish a sense of familiarity and identify themes within an extensive corpus that might be impractical to read entirely. This workflow hinges on the capability to generate topics from the data and subsequently delve into those topics. Such exploration necessitates the capacity to locate documents linked to a topic and realize the rationale behind their association with that topic. As topics produced without human intervention lack predefined names or inherent meanings, this exploration largely revolves around grasping the significance of each topic. We adopted an inductive or "bottom-up" approach [15] for the thematic analysis. This signifies that the identified themes are closely tied to the data itself and not influenced by the researcher's theoretical inclinations or interests in the subject matter. Inductive thematic analysis involves coding the data without attempting to impose it into a pre-established coding framework or the researcher's analytic assumptions. The whole process is presented in Figure 2. There are three distinct data sources—scientific publications, news blogs, and social media—that yielded three distinct collections of documents. Following data collection, we applied three parallel computational workflows for the text preprocessing, topic modeling, and visualizations. Subsequently, we engaged in a recursive process to interpret, validate, and label the derived topics. Finally, we integrated our findings and developed a thematic map.

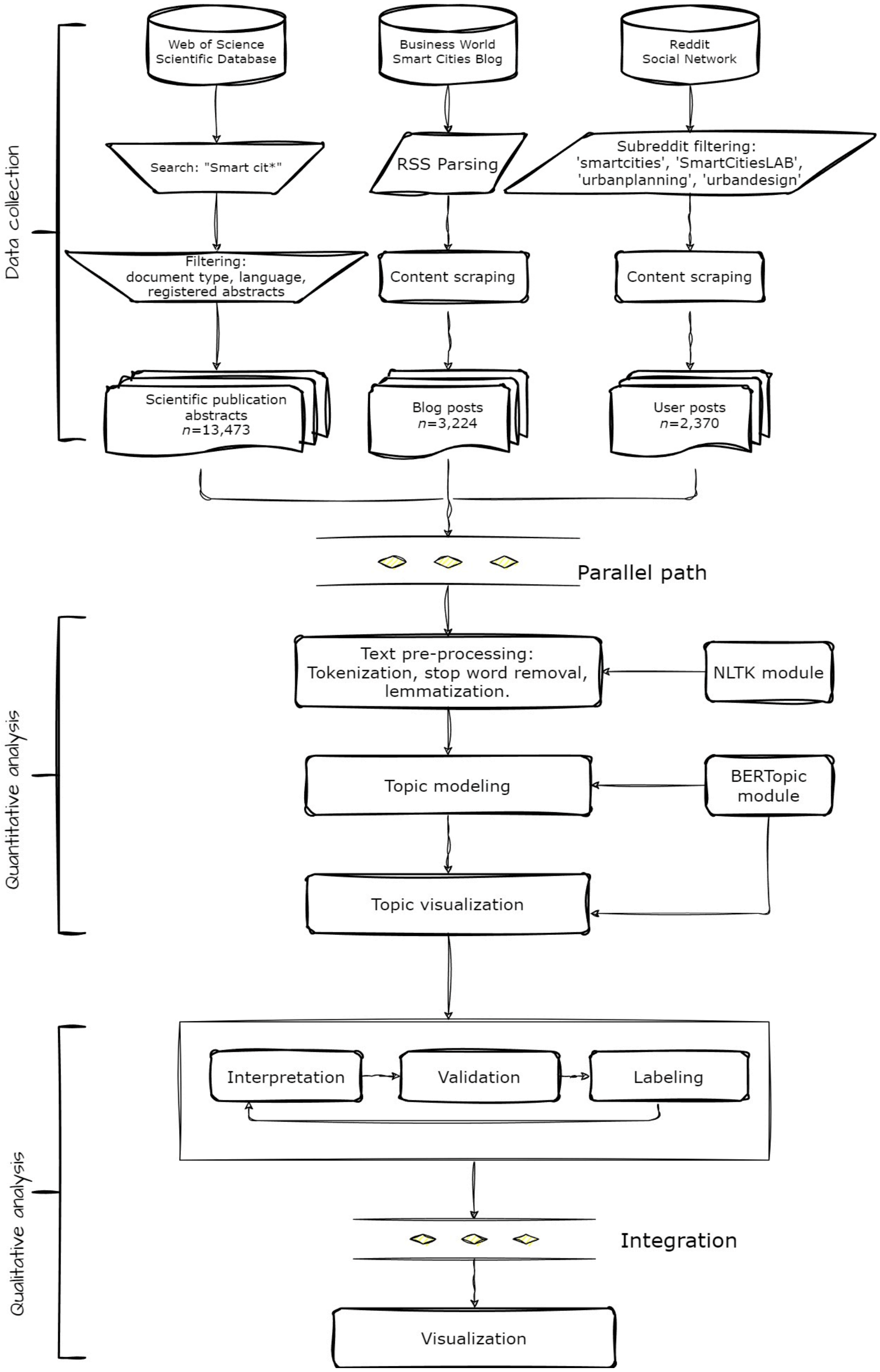
Data collection
Web of Science Scientific Database
Business World Smart Cities Blog
Reddit Social Network
Search: "Smart cit*"
RSS Parsing
Subreddit filtering: 'smartcities', 'SmartCitiesLAB', 'urbanplanning', 'urbandesign'
Filtering: document type, language, registered abstracts
Content scraping
Content scraping
Scientific publication abstracts n=13,473
Blog posts n=3,224
User posts n=2,370
Parallel path
Quantitative analysis
Text pre-processing: Tokenization, stop word removal, lemmatization.
NLTK module
Topic modeling
BERTopic module
Topic visualization
Qualitative analysis
Interpretation
Validation
Labeling
Integration
Visualization

Figure 2. The research workflow diagram.

2.2. Coding Language and Tools

The code has been implemented in Python 3.9 using Spyder IDE and Google Colaboratory as the execution environments and uses the Python modules listed in Table 1.

Table 1. Python modules utilized in this study.

Module	Use	URL	License
BERTopic	Topic modeling and visualizations	https://maartengr.github.io/BERTopic/index.html accessed on 30 July 2023	MIT
NLTK	Natural language toolkit	https://www.nltk.org/ accessed on 30 July 2023	Apache 2.0
Feedparser	RSS feed parsing	https://github.com/kurtmckee/feedparser/ accessed on 30 July 2023	BSD-2
Requests	HTTP requests	https://requests.readthedocs.io/en/latest/ accessed on 30 July 2023	Apache 2.0
BeautifulSoup	Web scraping	https://www.crummy.com/software/BeautifulSoup/ accessed on 30 July 2023	MIT
Asyncpraw	Reddit wrapper	https://asyncpraw.readthedocs.io/en/stable/ accessed on 30 July 2023	Simplified BSD
Matplotlib	Visualization library	https://matplotlib.org/ accessed on 30 July 2023	BSD
Pandas	Data analysis and manipulation	https://pandas.pydata.org/ accessed on 30 July 2023	BSD-3

2.3. Data Collection

Text data were collected from three distinct sources: (a) Scientific publication abstracts obtained from the Web of Science (WoS) database (https://www.webofscience.com/ accessed on 30 July 2023), (b) news blog posts from Businessworld (BW) Smart Cities blog (https://bwsmartcities.businessworld.in/ accessed on 30 July 2023), and (c) posts from the Reddit social network (https://www.reddit.com/ accessed on 30 July 2023).

2.3.1. Scientific Papers

To empirically support our proposed framework, this study acquired a corpus of scientific publication abstracts focused on the field of smart cities. These abstracts were sourced from the WoS database by conducting a search query formatted as "'smart cit*' (Title) OR 'smart cit*' (author keywords)", accounting for various endings like -y, -ies, -itizen, or -itizens, for the time span from 1999 to 2023, as of July 2023. The search yielded 14,980 results from its core collection. Initially, we refined the results by document types and retained only articles, proceedings papers, book chapters, and review articles. Subsequently, we focused on documents in the English language, leaving 13,810 documents. As some of the documents lacked registered abstracts in the database, we ended up with 13,473 abstracts for analysis. The gathered data offers valuable insights into the changing patterns of smart city research. Figure 3 represents the recent rise in interest in the field of smart cities. Notably, research in this domain witnessed significant acceleration, starting from the beginning of the 2010s. However, there was a period of stabilization in smart city research from 2018 to the present day. Despite these variations, smart city research continues to be of great interest to the scientific community.

2.3.2. News Articles

News articles possess sentence structures that facilitate the semantic detection process, particularly when leveraging computational methods. Such articles also enable the exploration of information and the acquisition of unforeseen perspectives, a capability that would be unattainable through predetermined survey questions [31]. BW stands as one of India's long-standing and highly regarded business media brands, providing an integrated media platform to effectively engage with a diverse audience through multiple channels and formats. The BW Smart Cities blog serves as an initiative aimed at charting and establishing connections among essential stakeholders, pivotal to the actualization of this vision. The blog maintains an RSS feed channel that encompasses all articles published since January 2015. Up until August 2023, a total of 3224 news articles on smart cities have been published. We extracted the textual content using the Python module Feedparser (https://pypi.org/project/feedparser/ accessed on 30 July 2023) and performed scraping through the employment of the BeautifulSoup (https://pypi.org/project/beautifulsoup4/ accessed on 30 July 2023) module.

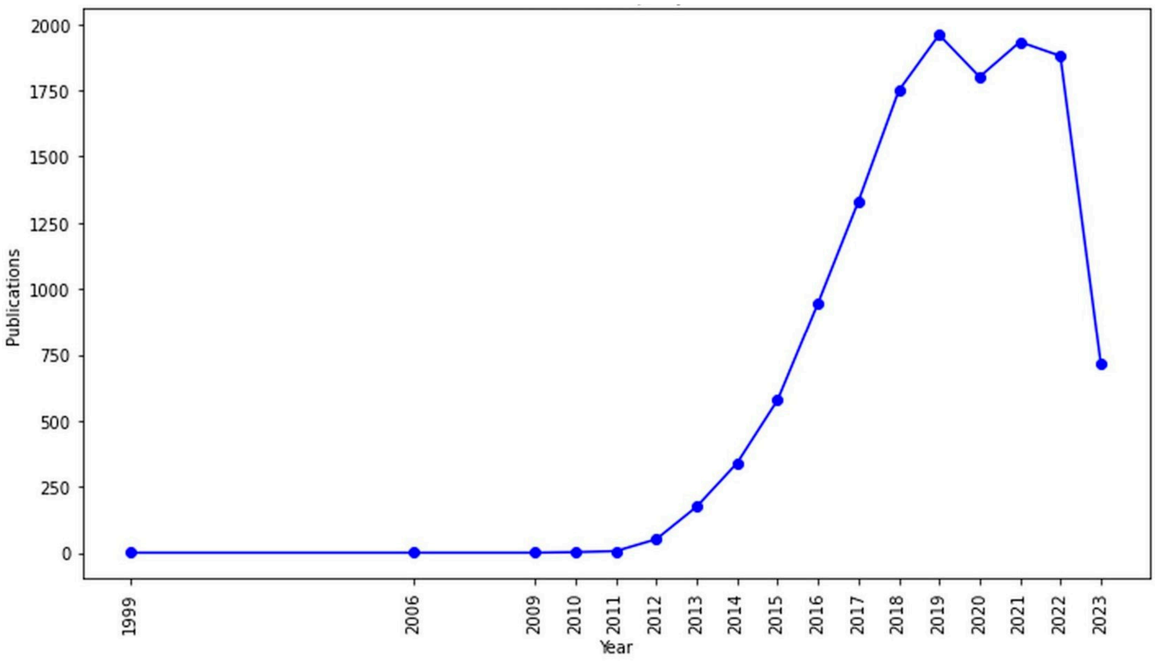

Figure 3. Evolving trends on smart city research based on WoS data.

2.3.3. Social Media

Nowadays, people around the world use social media as a means of communication, establishing connections, and engaging with other users [32]. This has led to a rapid generation of substantial amounts of big data [33], prompting a transformative shift in the urban civic management of cities. This shift entails a movement from traditional top-down hierarchical city planning and civic management, driven by governmental agencies, to a more participatory approach supported by citizen engagement in urban planning and civic management efforts [27]. Analyzing how people perceive urban spaces is crucial for all stakeholders involved in city development, including policymakers, local organizations, and urban designers [31], to deliver to citizens novel or improved services [34].

For our study, we chose Reddit, which is a widely used social media platform where users from around the world submit, vote, and comment within over 100,000 active interest-based communities known as subreddits. For our analysis, we examined 2370 posts, extracted using the Python module called Asyncpraw (https://asyncpraw.readthedocs.io/ accessed on 30 July 2023) from four subreddits ('smartcities', 'SmartCitiesLAB', 'urbanplanning', 'urbandesign') that are relevant to smart city planning.

2.4. Text Data Preprocessing

Once the data had been collected from the three distinct sources, we proceeded with the necessary preprocessing steps for each dataset. In most cases, data preprocessing is a crucial prerequisite for achieving accurate outcomes, and the adage "garbage in—garbage out" certainly holds true [35]. This saying implies that if our data are poorly formatted, it is likely that our results will be of poor quality [30].

Data preprocessing involves performing fundamental transformations to make the data more conducive to subsequent and meaningful analysis. The initial preprocessing step involves text normalization, which entails converting all letters to lowercase. The second preprocessing step entails removing stop words, which are the most common words in a language [36]. These words typically lack significant meaning and are usually excluded from texts. The third and final preprocessing procedure we conducted was reducing the size of the document–term matrix, which helps in enhancing computing efficiency and achieving improved results through lemmatization, a powerful technique. Lemmatization involves morphological analysis to identify the root word [30], reducing all words to their base "lemma" using a lexicon and conjugation rules.

2.5. Topic Modeling

Topic modeling is an automated, unsupervised technique that discovers the underlying topics within a collection of documents, along with the connections between each document and these topics [37]. The algorithms can achieve this with minimal human involvement, rendering the approach more deductive compared to the conventional techniques used in analyzing social and human science texts. While not flawless, when employed thoughtfully and executed with caution, the method appears to consistently produce plausible interpretations of the texts [38]. The output of topic modeling comprises two elements: the proportions of words within topics and the proportions of topics within documents. Upon completion of the training process, both output elements can be extracted from the model and subjected to human analysis [16]. Since topic modeling portrays each document as an intricate amalgamation of various topics and each topic as an intricate amalgamation of multiple words, it is also employed as a method to categorize documents according to the outcomes of inferred topics [34].

Ogunleye et al. [39] traced the evolution of topic modeling methods, dating back to 1990 when Deerwester et al. [40] introduced latent semantic indexing (LSI). LSI employs singular value decomposition (SVD) on a large term–document matrix to identify a linear subspace that captures the connection between words and documents. Nevertheless, LSI has limitations, as it does not assign probabilities to topics. Addressing this, Hofmann [41] developed probabilistic latent semantic indexing (pLSI), which also received criticism for its inability to account for the generative probabilistic models of documents. Consequently, LDA was conceived by Blei et al. [42] as an advancement. Nevertheless, these conventional topic modeling methods have faced backlash due to their inability to manage data scarcity, a challenge particularly noticeable in concise texts. However, considerably more recent and advanced techniques like BERTopic [43] are swiftly gaining traction, especially in applications involving social media [11]. Unlike conventional topic models that depend on the "Bag-of-Words" representations, which disregard the inherent semantic connections among words by employing a mere lexical bag format, BERTopic, functioning as a neural topic model, has the capability to depict words as multidimensional vectors. This allows it to capture contextual information, resulting in more precise and comprehensive attributes [44].

Expanding on the reasons provided above, we intend to employ the BERTopic model for the more efficient and accurate extraction of smart city aspects.

BERTopic

BERTopic was introduced by Grootendorst [43] as "a topic model that leverages clustering techniques and a class-based variation in term frequency–inverse document frequency (TF-IDF) to generate coherent topic representations". BERTopic demonstrated impressive performance in terms of two widely used validation measures, topic coherence (TC) and topic diversity (TD), in comparison to LDA, NMF, CTM, and Top2Vec in three different benchmark datasets [45]. Furthermore, this approach demands less effort in hyperparameter tuning and does not necessitate a predetermined number of topics, providing BERTopic an advantage over LDA [44]. Thakur et al. [46] underscored two additional advantages of BERTopic in comparison to the alternative methods: its straightforward, out-of-the-box usability and its innovative, interactive visualization capabilities. BERTopic employs a modular architecture comprising five core sequential functions, illustrated in Figure 4.

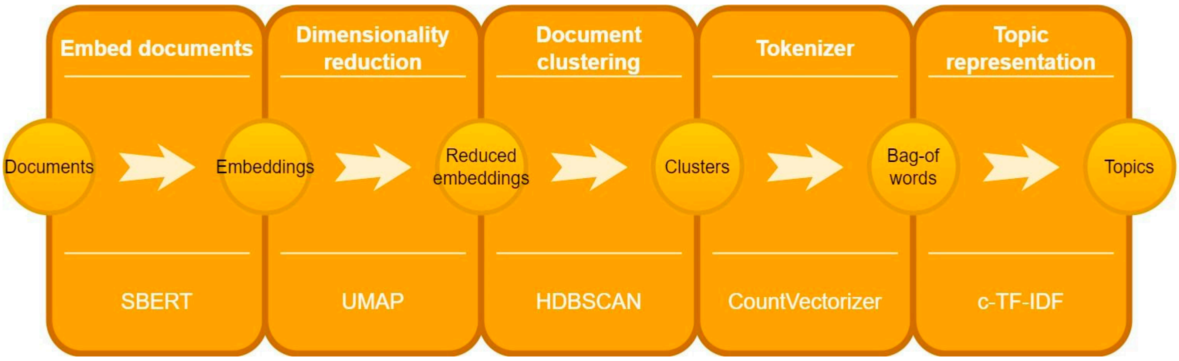

Figure 4. The five steps of BERTopic.

More specifically, the model:

1. Converts a document to numerical representations in a vector space, known as embeddings, using a pre-trained language model. This process involves obtaining information at the document level. To fulfill this function, we employed the SBERT sentence transformer [47], which is a modified version of the pre-trained Bidirectional Encoder Representations from Transformers (BERT) network [48]. SBERT utilizes siamese and triplet network structures to generate semantically meaningful sentence embeddings. For our study, we opted for the 'all-MiniLM-L6-v2' pre-trained sentence transformer model included in SBERT. This model, trained for English language semantic similarity tasks, is well-suited for a wide range of use cases [43].

2. Reduces the dimensionality of document embeddings. Due to the "curse of dimensionality", clustering algorithms struggle to efficiently handle high-dimensional data. Since SBERT maps sentences to a 384-dimensional dense vector space, there arises a requirement for reducing the dimensionality of these representations. For our use case, we selected the uniform manifold approximation and projection (UMAP) [49], a practical and scalable algorithm suitable for dimensionality reduction in real-world data. UMAP is a technique capable of retaining aspects of a dataset's local and global structure while reducing its dimensionality.

3. Creates clusters of semantically similar documents, each representing a distinct topic. Following the reduction in embeddings, the subsequent step involves generating clusters of documents that share semantic similarity. BERTopic incorporates a variation in widely used clustering algorithms, such as K-means or agglomerative clustering. However, as these techniques are susceptible to noise, we employed the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) [50]. HDBSCAN is adept at identifying clusters of various shapes and has the valuable capability of detecting potential outliers. Consequently, we avoid the imposition of documents into clusters where they may not indeed belong [43].

4. Tokenizes each cluster, generating a bag-of-words representation. The selection of HDBSCAN as the chosen clustering algorithm acknowledges the potential for clusters to manifest diverse degrees of density and shapes. In essence, we seek a topic representation methodology that abstains from presupposing any a priori assumptions concerning the inherent structural compositions of clusters. To this end, we consolidated all documents within a cluster into a single, albeit lengthy, document. This composite document functions as the embodiment of the entire cluster. Importantly, this bag-of-words representation operates on the cluster level, distinct from the document level. This distinction is crucial because our interest lies in the words at the topic level or, more precisely, the cluster level.

5. Weights tokens using a class-based variant of TF-IDF to extract the topic representation from each cluster. The employment of the class-based term frequency–inverse document frequency (c-TF-IDF) serves as a metric for gauging the significance of each word concerning specific topics and for identifying the most representative word associated with each topic. The outcome offers an estimation of word importance within a given

cluster. This technique seamlessly lends itself to the generation of matrices for document-topic distribution and topic-word distribution, thereby proving conducive to the matrix generation process [44].

2.6. Evaluation, Visualization, Interpretation, and Optimization

Upon establishing our topic model, we were empowered to examine our corpus and gain deeper insights into the inherent characteristics of our topic models. Post-processing activities related to topic models encompass activities like model evaluation, pinpointing the most representative documents for each topic, visualization, and the interpretation of topics, as well as optimization of the model.

2.6.1. Evaluation

When the computations for topic modeling are completed, it becomes necessary to validate the outcomes. Cai et al. [51] enumerates various reasons for the potential appearance of suspiciousness in extracted topics. These include (1) the merging of two or more distinct themes into a singular theme; (2) the extraction of two themes that, according to human perception, appear as duplicates; (3) the existence of topic keywords extracted without apparent coherence; (4) topics encompassing an excessive number of generic terms; (5) the extraction of topics based on seemingly unrelated terms; (6) inconsistency between the extracted topics and human judgment; (7) the perception of extracted topics as irrelevant; (8) the absence of a clear correlation between the topics and associated documents, and (9) the extraction of multiple closely similar topics. Evaluating the topic models can often prove to be a challenging activity, as the interpretability of the outcomes is not consistently assured with unsupervised methodologies [35]. This culminates in the presentation of a list of words ranked by relevance, and assigning a label to the topic based on these words can indeed be intricate.

Asmussen and Møller [19] delineated three approaches for validating the results of a topic model: statistical, semantic, or predictive. Statistical validation entails the application of statistical methods, such as topic perplexity and coherence [42] or Cohen's kappa [52], to test quantitatively the model's underlying assumptions. Semantic validation involves comparing the topic model results with expert reasoning, ensuring that the outcomes align with semantic coherence. Put differently; it questions whether the clustering of documents into topics is conceptually logical and ideally verified by an expert. For instance, a method could involve the manual coding of papers and subsequent comparison of the coding to the topic model's output. Predictive validation is employed when an external event can be correlated with an incident not explicitly present in the papers.

In our case, we opted for semantic validation through a two-step process. Initially, as Srinivasa-Desikan [30] suggests, we leveraged the visualization capabilities provided by the BERTopic module to gain a holistic understanding of the underlying latent themes. Subsequently, we conducted an in-depth examination of the most representative documents within each topic to corroborate our initial perceptions.

2.6.2. Visualization

Visualization can offer comprehensive insights into the outcomes of a topic model and aid in identifying modeling challenges. Kherwa and Bansal [53] argue that the most common way to understand topics is through visualization, and in this view, BERTopic visualization capabilities directly address the challenge of model understanding by providing several useful tools to the researcher. These tools support the user in exploring the corpus, moving between levels of granularity like terms, topics, documents, and the hierarchy of the model.

The visualization capabilities of BERTopic are intended to assist users in interpreting the topics within a fitted topic model applied to a collection of documents. The module draws information from a fitted topic model to generate interactive web-based visualizations. These visualizations are designed for interactions, as they can be saved as standalone

HTML files, facilitating effortless sharing. Users can concentrate on distinct sections within these interactive visualizations to enhance their comprehension of model intricacies. By employing a drag-and-drop functionality, users can select a cluster of topics, thereby unveiling more detailed information about the chosen cluster.

2.6.3. Interpretation

As Gillies et al. [11] stated, there exists an intriguing analogy between the objectives of topic modeling and qualitative research. Nevertheless, there are some differences between them. The automated nature of topic models renders them well-suited for extensive datasets; however, this characteristic also results in a relative deficiency when compared to the nuanced human interpretation inherent in qualitative research. Within the context of topic modeling, human interpretation often centers on the final stage—the interpretation of the algorithmic output.

The purpose of post-processing is to identify and assign labels to the topics derived from the BERTopic model. The algorithm generates topics that encapsulate clusters of words often found co-occurring in the documents, along with a compilation of the most representative documents for each topic. However, labeling topics is not a straightforward endeavor. A specific challenge in characterizing topics is to determine a concise, descriptive name for each topic. Since topics are treated as latent variables representing underlying themes in the data, topic models learn topics that lack inherent canonical descriptions. The process of labeling topics involves identifying the principal topic of each topic group. Naturally, this is a subjective issue, potentially resulting in varying labels for topics contingent on the researcher [19]. To mitigate the possibility of misidentified topics, it is necessary to conduct a thorough review of the most common words associated with each topic and closely examine the most representative documents.

2.6.4. Optimization

After reviewing and labeling the identified topics to ensure their alignment with our comprehension of the smart city concept, the subsequent step involves refining the model. BERTopic provides a bunch of options for refining our model, including manual or automatic topic reduction, updating topic representation, incorporating custom labels, reducing outliers, as well as hyperparameter tuning options for each module used in our model, such as adjusting the number of topics generated or fine-tuning the preprocessing steps. Additionally, since we intend to integrate outcomes from three models generated from separate sources, we must tackle challenges such as merging distinct themes into singular themes, identifying duplicates, finding topics laden with an excessive number of generic terms or unrelated terms, and addressing the absence of a clear correlation between topics and associated documents.

2.7. Thematic Analysis

Thematic analysis is a more challenging process than simply counting words in a document, focusing instead on recognizing and elucidating implicit and explicit concepts within the data, referred to as themes [54]. A theme encapsulates an important aspect of the data concerning the research question and signifies a degree of patterned response or significance within the dataset. A pivotal query, in terms of coding, pertains to determining what qualifies as a pattern or theme or the requisite 'extent' of a theme. This matter revolves around prevalence, encompassing the spatial prominence within each data item and prevalence across the entirety of the dataset. Ideally, multiple instances of the theme should be present throughout the dataset, although an increased number of instances does not necessarily amplify the theme's significance. Given the qualitative nature of this analysis, there is no definitive answer regarding the proportion of the dataset that must exhibit evidence of a theme to identify it as such [15]. Our analysis encompasses the integration of the information gleaned from the three topic models, followed by the visualization of the themes through a thematic map, to unveil the conceptual structure of the smart city paradigm.

3. Results

We conducted topic modeling analyses on three separate datasets: (a) abstracts from scientific publications, (b) posts from a news blog, and (c) entries from a social media platform. The model generated 145, 52, and 41 topics, respectively. The comprehensive results are presented in Tables A1, A2 and A4, included in Appendix A.

3.1. Scientific Publication Abstracts

The initial step in refining the topics of the scientific publication abstracts is to examine the intertopic distance map (Figure 5) and the hierarchical clustering (HC) dendrogram (Figure 6). These visualizations offer a panoramic perspective of the topics whilst enabling a thorough examination of the terms most strongly associated with each individual topic. To ensure a more comprehensive analysis, interactive visualizations are available in Supplementary Materials section. The intertopic distance map depicts each topic as a circle in the two-dimensional plane, with their centers positioned based on the calculated intertopic distances. The positioning is achieved through multidimensional scaling, which projects the intertopic distances into two dimensions. The visualization in the two dimensions is facilitated by employing Python's Plotly module (https://plotly.com/ accessed on 30 July 2023), enabling the creation of an interactive display [43]. The HC dendrogram portrays topic similarity at the document level. This visualization supports the comprehension of the intellectual structure of topics within the domain of smart cities.

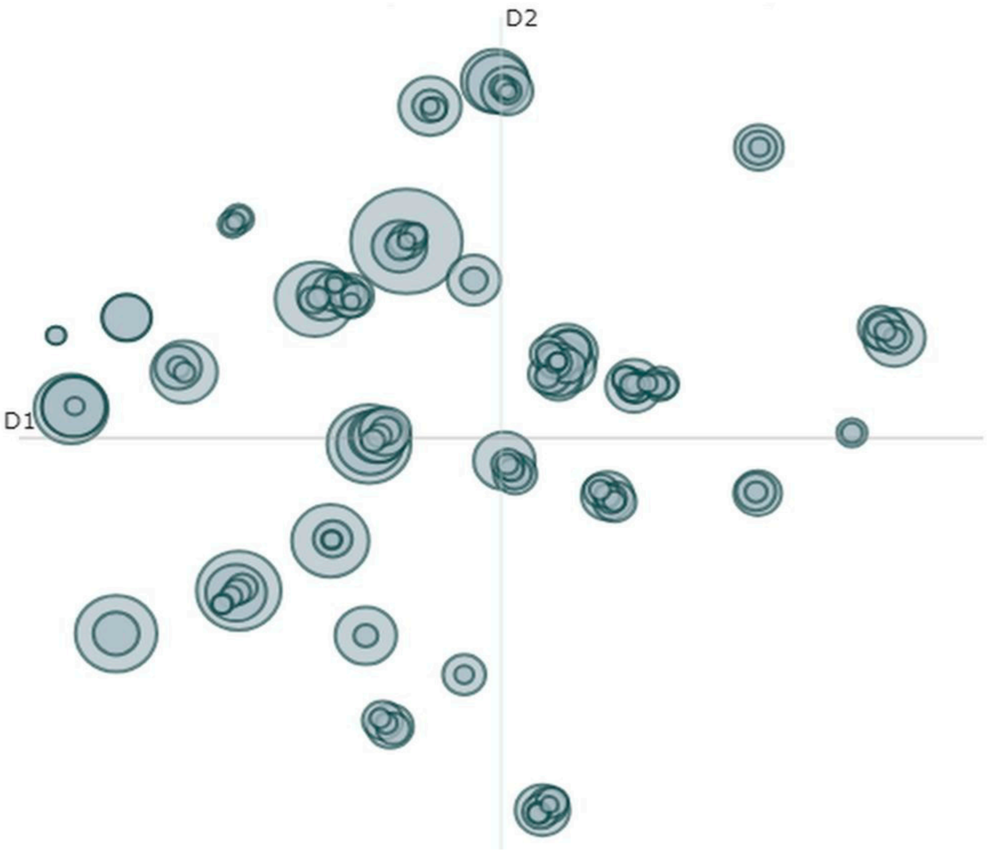

Figure 5. Intertopic distance map for scientific publication abstracts.

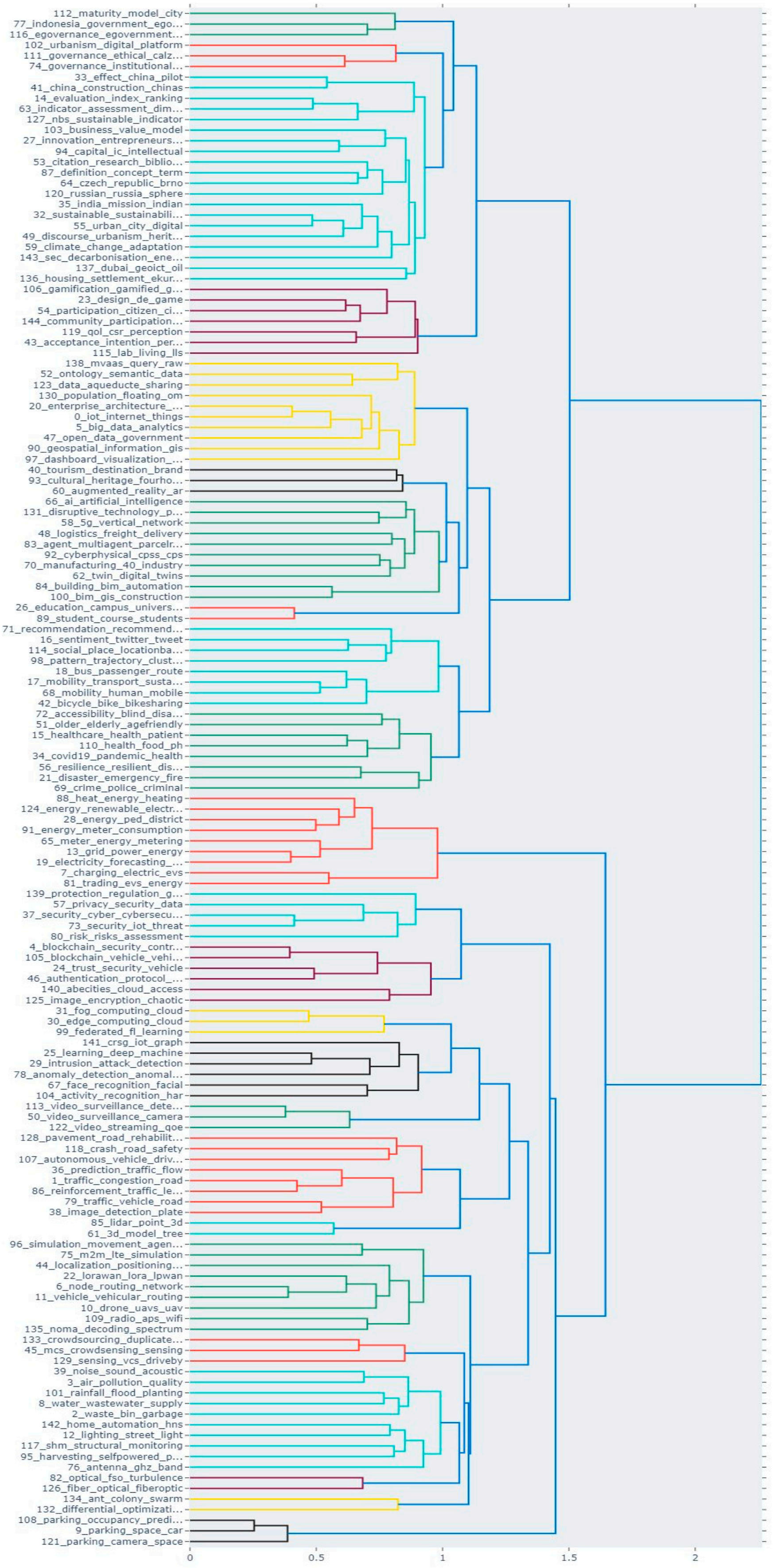

Figure 6. Dendrogram for scientific publication abstracts.

The analytical procedure we employed facilitates the discernment of topics characterized by their concrete validity whilst enabling the identification of topics that need to be refined, merged, or excluded. For these achievements, when our choices were not trivial and needed more examination, we thoroughly read the more representative abstracts and, where available, even delved into complete textual content.

After our analysis, we proceeded to merge the topics, as shown in the following list: {{0, 141}, {1, 17, 36, 79, 86, 105}, {6, 22, 75, 95, 109}, {7, 81}, {8, 101, 134}, {9, 108, 121}, {11, 24}, {13, 19, 65}, {14, 63, 112, 127}, {15, 34, 104}, {16, 98, 114}, {18, 68}, {21, 56}, {23, 60, 106}, {26, 89}, {28, 88, 91, 124}, {29, 67, 78, 104}, {30, 31}, {33, 143}, {37, 73, 80, 140}, {38, 125}, {40, 93}, {45, 129, 133}, {48, 83}, {50, 113, 122}, {52, 138}, {53, 87, 119}, {54, 144}, {55, 130, 131}, {57, 139}, {61, 62, 92, 96}, {74, 77, 111, 116}, {76, 135}, {82, 126}, {84, 100}, {85, 107}}. Topics 35, 41, 43, 49, 64, 99, 102, 120, 132, and 137 were considered as outliers. The outliers are identified in the topics related to specific places (countries or cities), such as China or Brno, and some others that do not seem to include words with a coherent meaning. Subsequently, we examined the refined model and labeled the resultant topics.

Figure 7 illustrates the labeled topics along with the c-TF-IDF scores of the most frequently occurring terms within each. Figure 8 represents the hierarchical structure of the refined topic model. Figure 9 portrays the hierarchical structure of the documents within each corresponding topic, supplied with embeddings. In the interactive visualization, the user can discern each document in the dataset associated with the relevant topic by simply hovering the mouse over it.

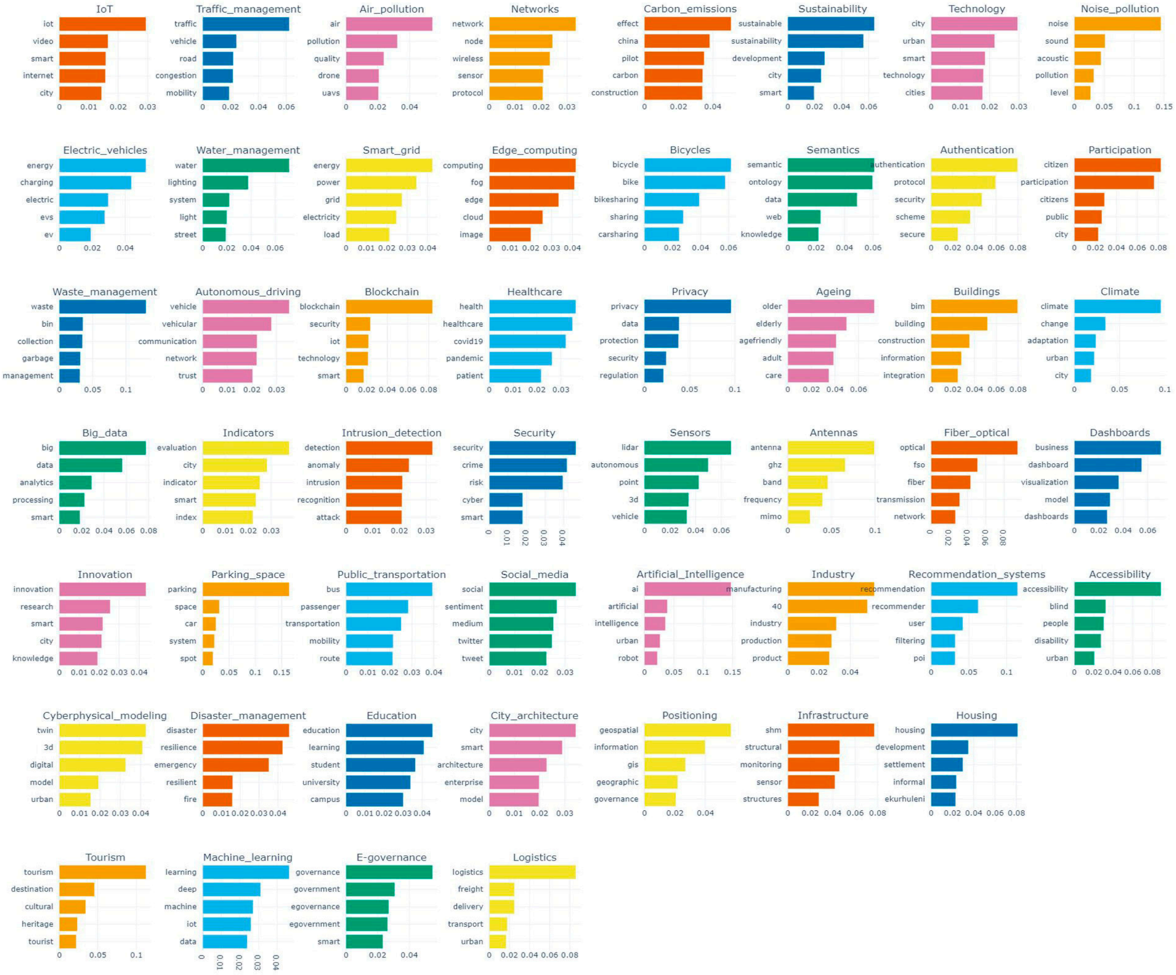

Figure 7. Topic-word scores for scientific publication abstracts.

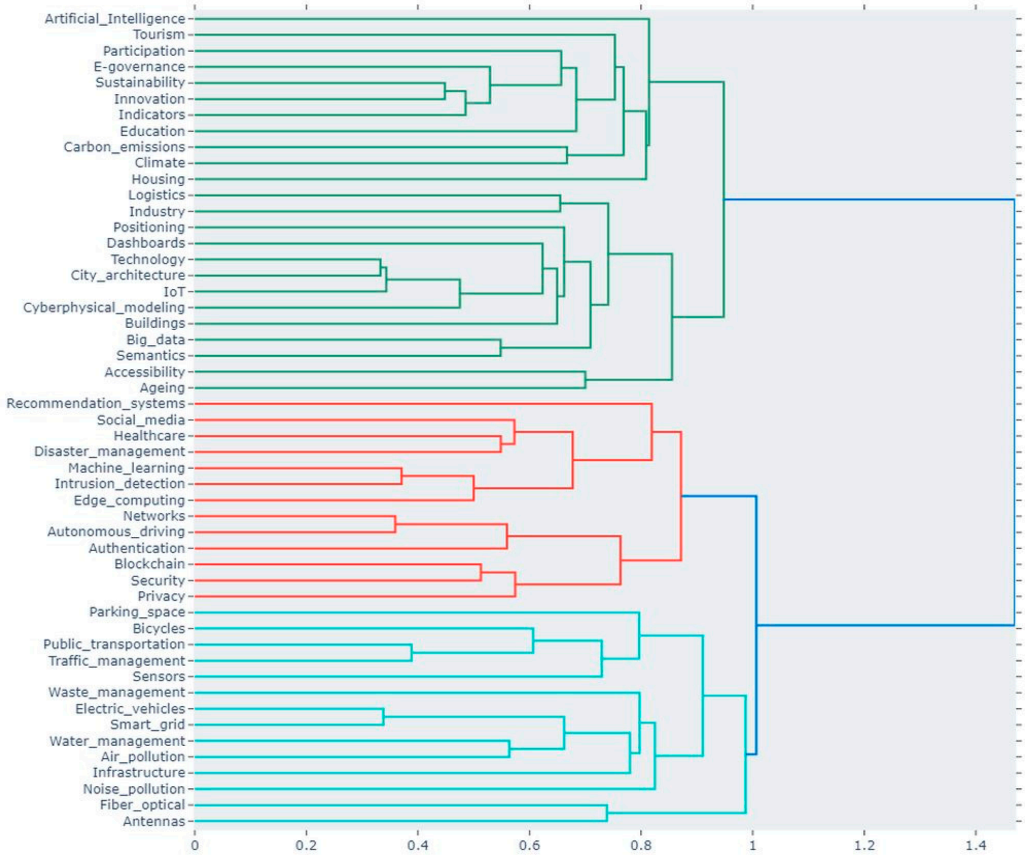

Figure 8. Dendrogram for the refined model of scientific publication abstracts.

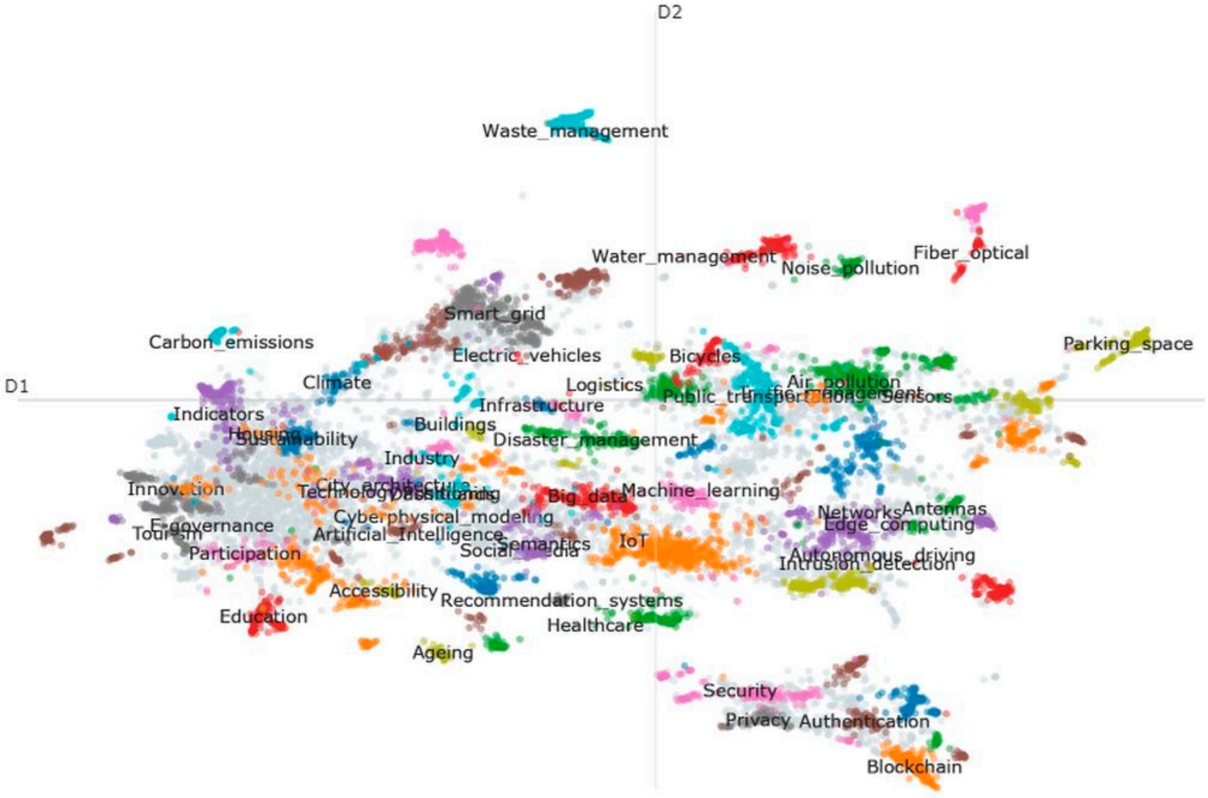

Figure 9. Hierarchical document structure of corresponding topics for scientific publication abstracts.

With the production of topic embeddings using both c-TF-IDF and embedding techniques, it becomes feasible to create a similarity matrix by directly employing cosine similarities on these topic embeddings. The outcome yields a matrix that signifies the degree of similarity between specific topics, illustrated in Figure 10 as a heatmap.

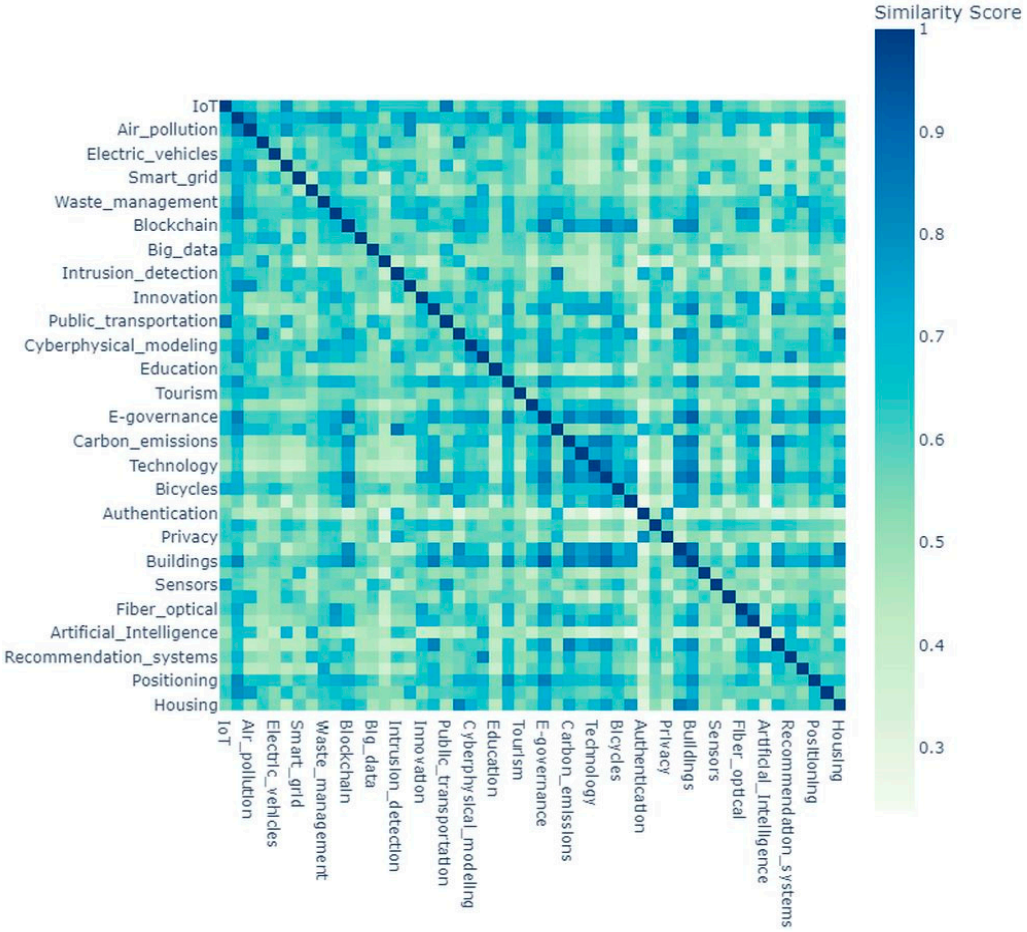

Figure 10. Similarity matrix for scientific publication abstracts.

3.2. News Blog Posts

For the news blog posts dataset, we followed the identical workflow as for the scientific publication abstracts dataset. Initially, we examined the intertopic distance map (Figure 11) and the HC dendrogram (Figure 12). Following that, we conducted a comprehensive review of the more representative blog posts. Upon completing our analysis, we advanced to merge the topics, as outlined in the following list: {{−1, 24, 25, 30, 33, 42,45, 47, 50}, {0, 32, 34}, {1, 51}, {4, 13, 13}, {8, 49}, {12, 22}, {19, 41}, {21, 39, 43, 46, 48}}. Topics 24, 25, 30, 33, 42, 45, 47, and 50 are considered as outliers.

Subsequently, we examined the refined model and labeled the resultant topics. Figure 13 illustrates the labeled topics along with the c-TF-IDF scores of the most frequently occurring words within each. Figure 14 represents the hierarchical structure of the refined topic model. Figure 15 portrays the hierarchical structure of the documents within each corresponding topic, supplied with embeddings. Figure 16 illustrates a heatmap displaying a matrix depicting the levels of similarity among distinct topics.

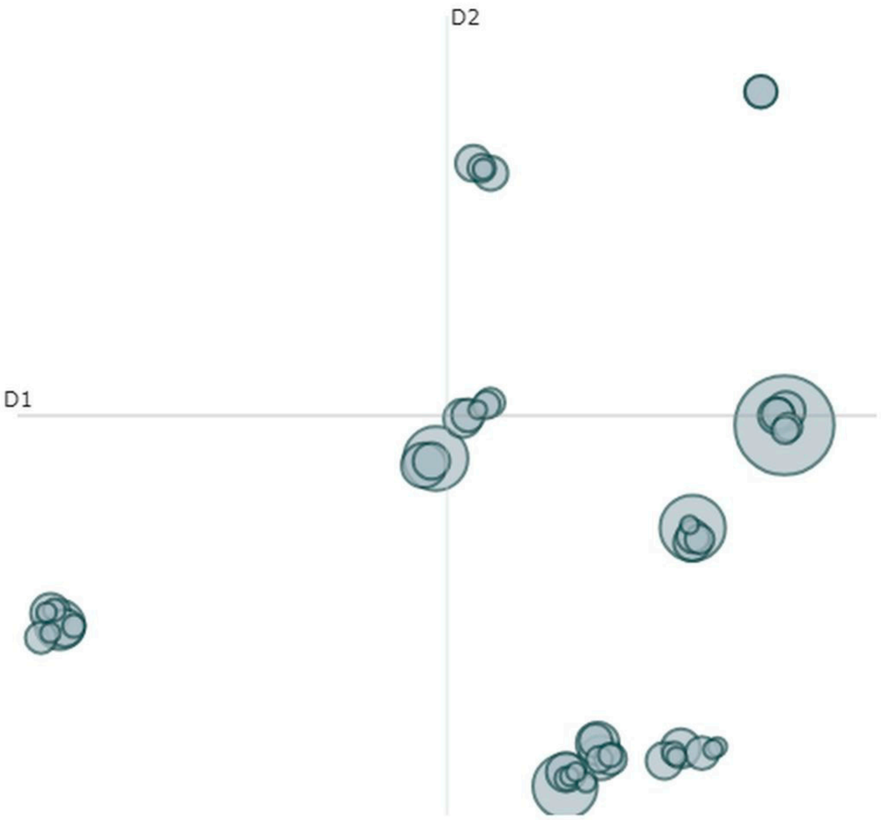

Figure 11. Intertopic distance map for news blog posts.

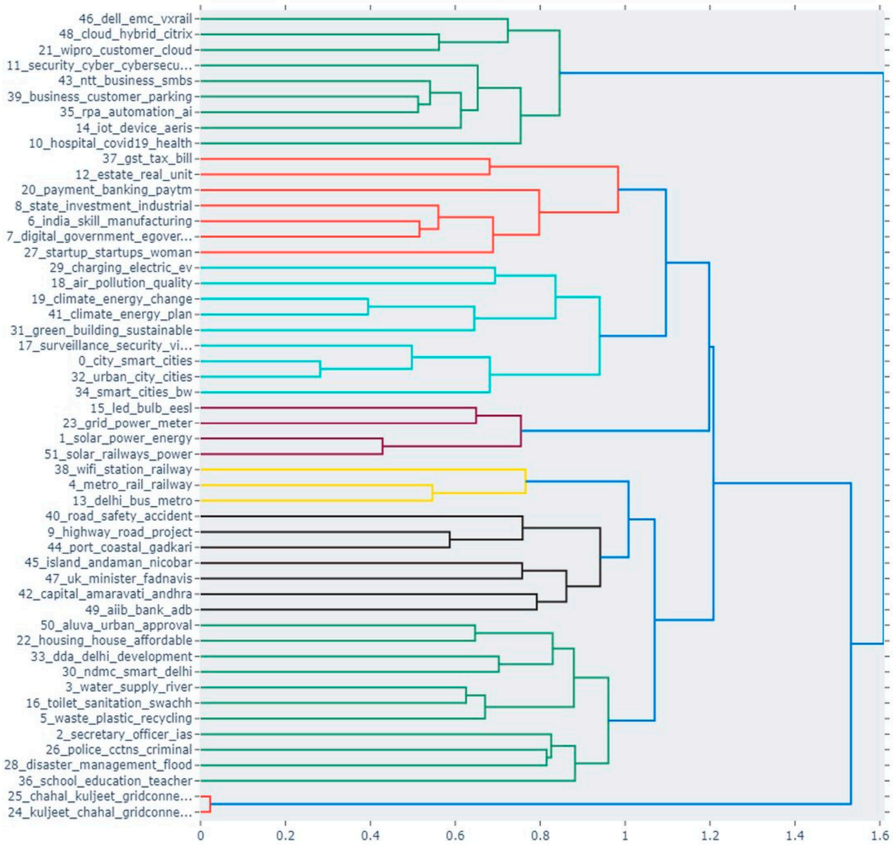

Figure 12. Dendrogram for news blog posts.

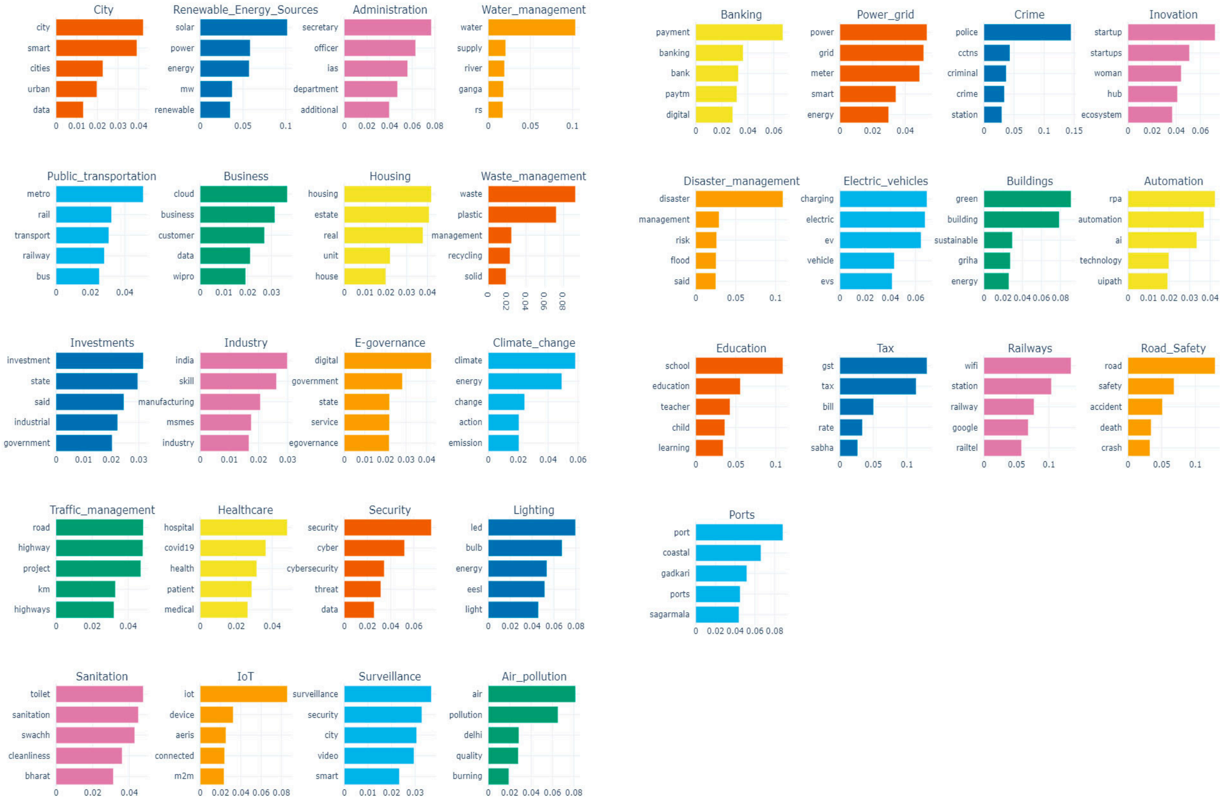

Figure 13. Topic word scores for news blog posts.

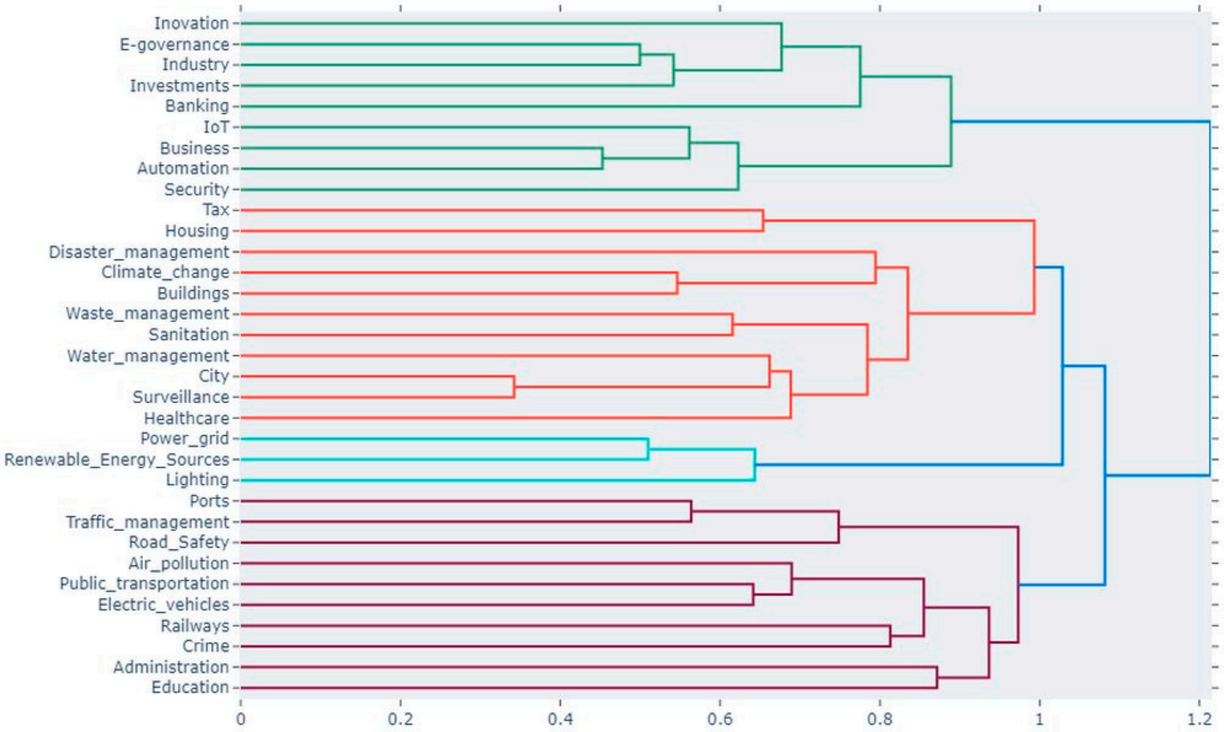

Figure 14. Dendrogram for the refined model of news blog posts.

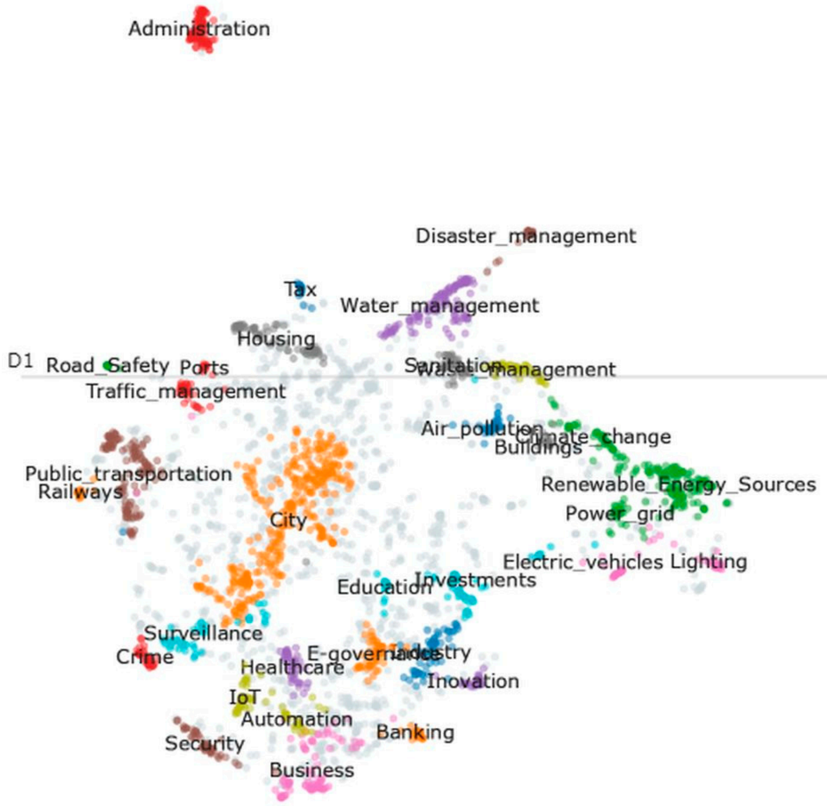

Figure 15. Hierarchical document structure of corresponding topics for news blog posts.

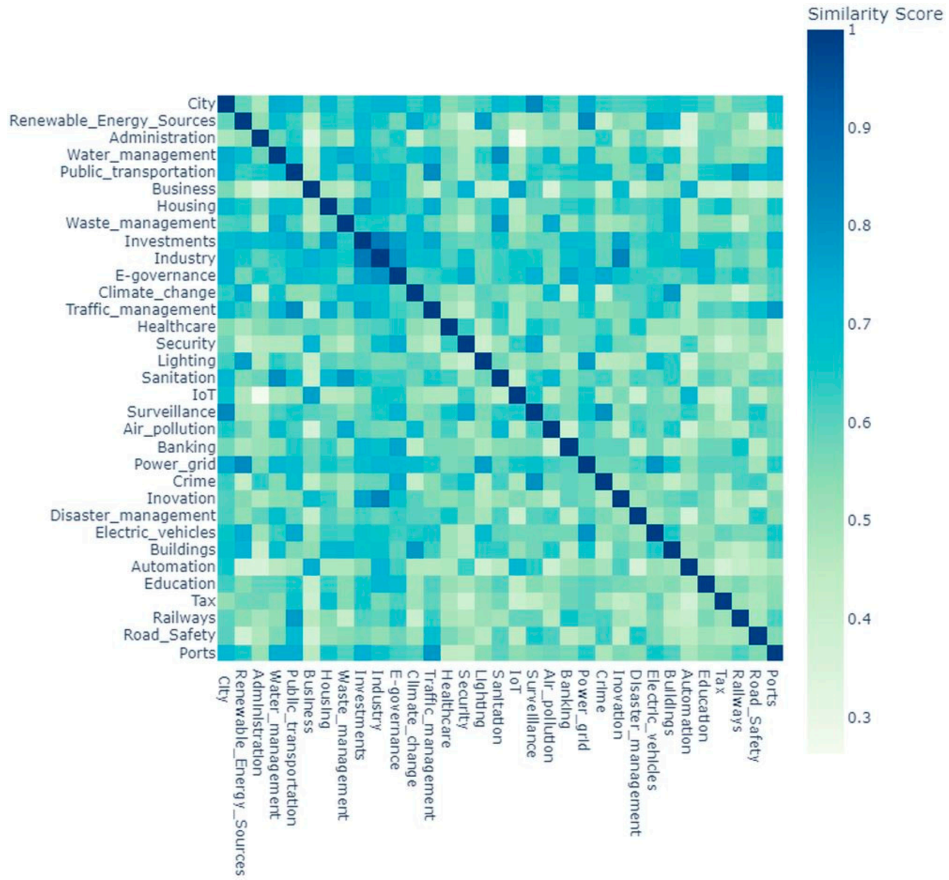

Figure 16. Similarity matrix for news blog posts.

3.3. Social Media Posts

For the social media entries dataset, we adhered to the same workflow as in the previous datasets. Consequently, we examined the intertopic distance map (Figure 17) and the HC dendrogram (Figure 18). Following this, we conducted a comprehensive review of the more representative Reddit posts. Upon the conclusion of our analysis, we proceeded to merge the topics, as outlined in the following list: {{0, 6}, {1, 4, 24, 34}, {2, 8, 23, 36}, {7, 10}, {11, 16}, {13, 35}, {14, 20}, {19, 22, 32}, {21, 37, 39}, {26, 29, 38}}. Topics 25, 28, 31, 33, and 40 were considered as outliers.

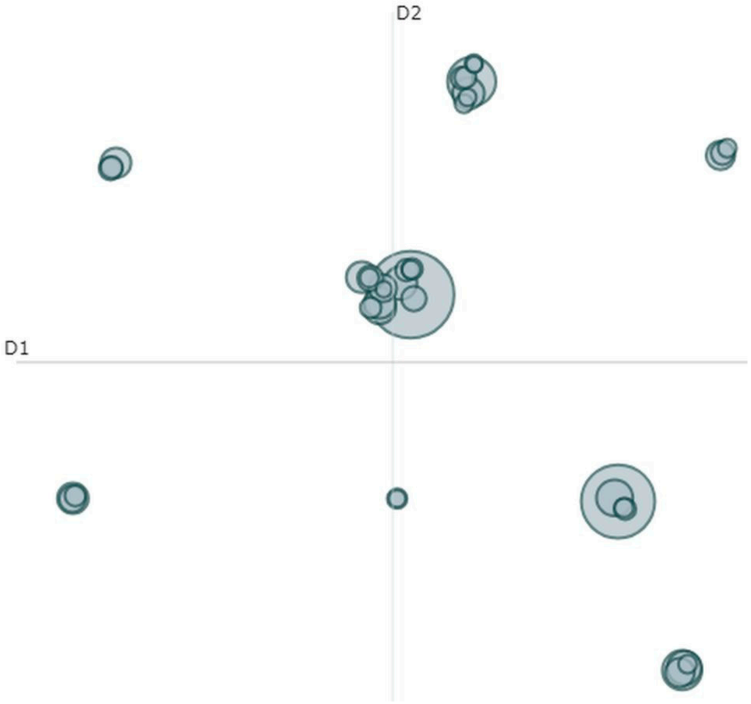

Figure 17. Intertopic distance map for social media posts.

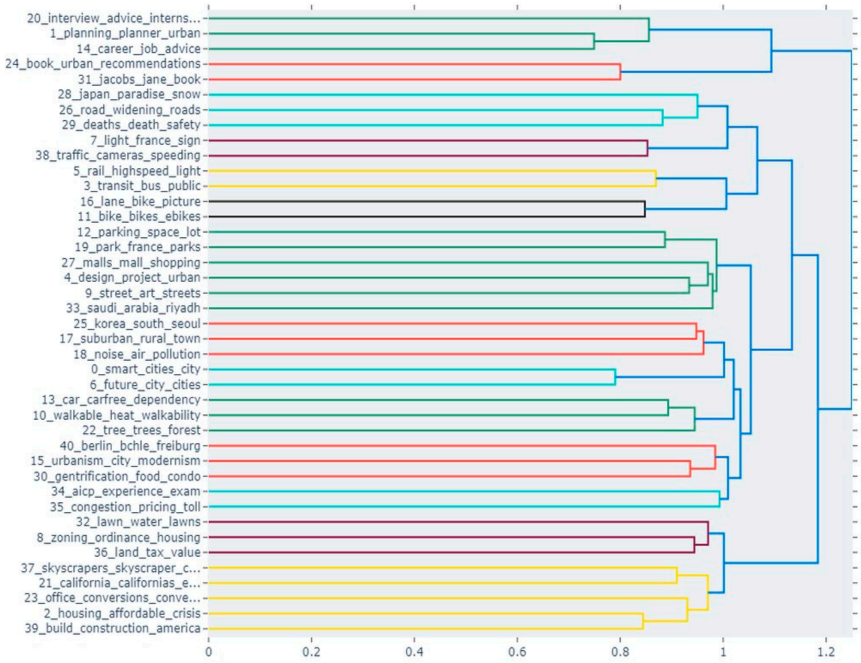

Figure 18. Dendrogram for social media posts.

Subsequently, we analyzed the refined model and labeled the resulting topics. Figure 19 illustrates the labeled topics along with the c-TF-IDF scores of the most frequently occurring words within each topic. Figure 20 depicts the hierarchical structure of the refined topic model. Figure 21 portrays the hierarchical arrangement of the documents within their corresponding topics, enriched with embeddings. Figure 22 showcases a heatmap that presents a matrix illustrating the degrees of similarity among distinct topics.

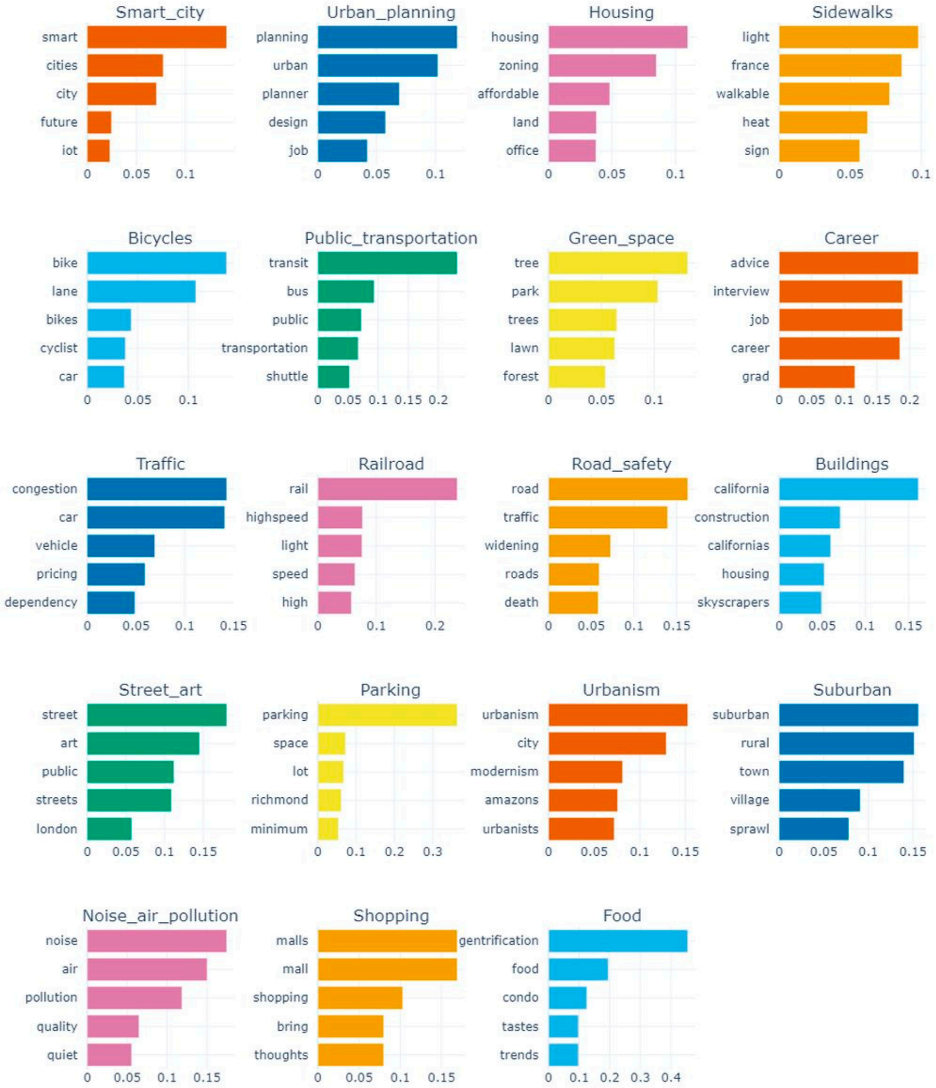

Figure 19. Topic word scores for social media posts.

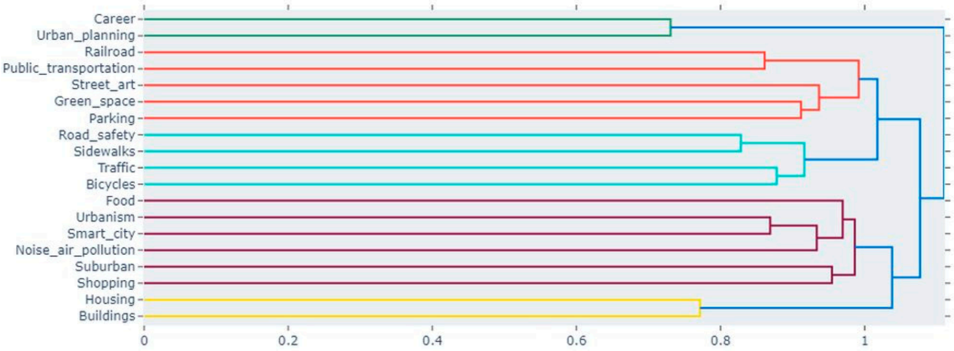

Figure 20. Dendrogram for the refined model of social media posts.

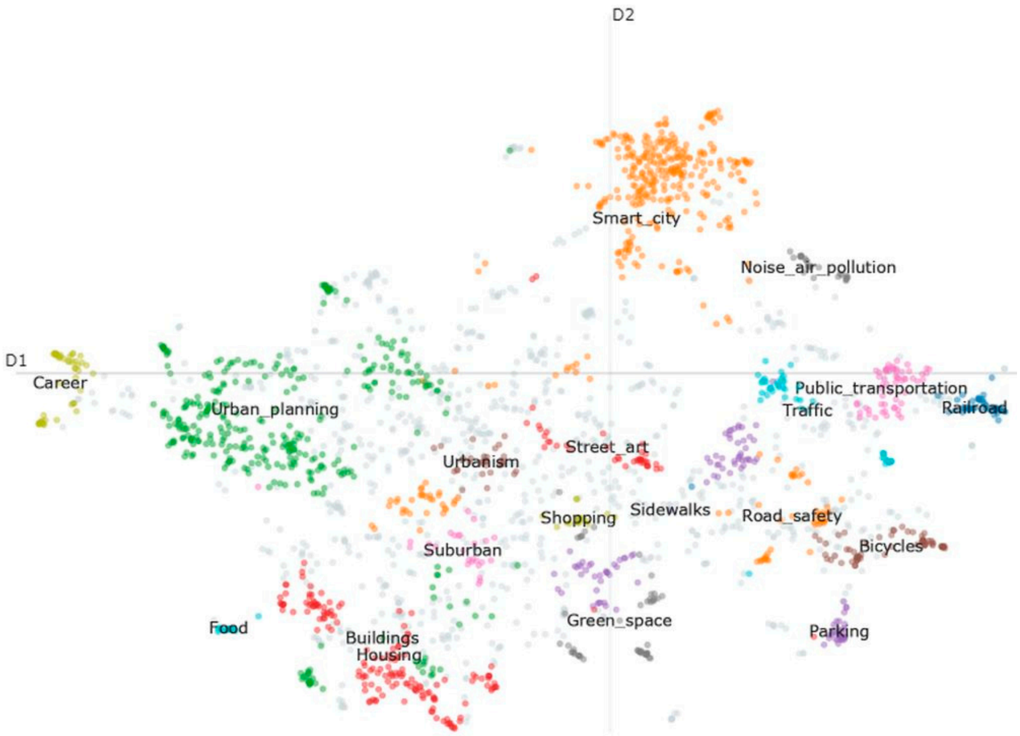

Figure 21. Hierarchical document structure of corresponding topics for social media posts.

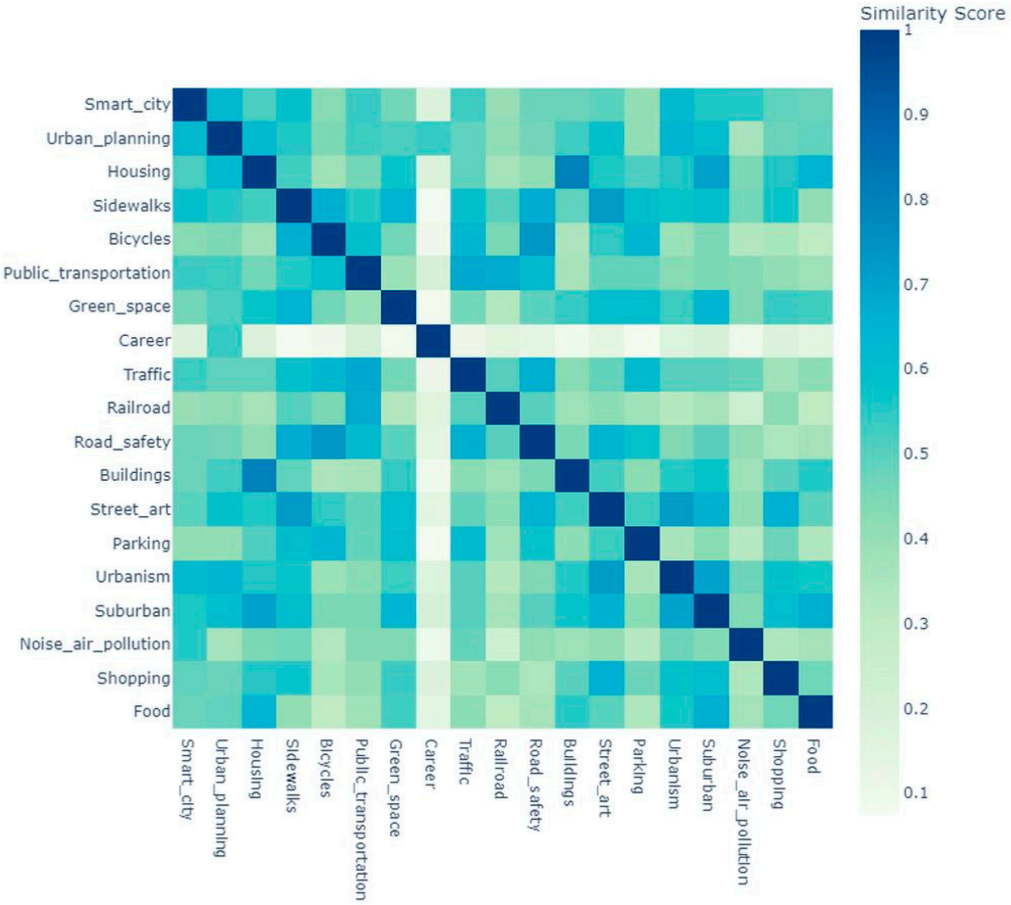

Figure 22. Similarity matrix for social media posts.

We observed the emergence of certain shared topics across all datasets, along with a variety of identical topics within each dataset, as summarized in Table 2.

Table 2. Common and identical topics in each dataset.

Common	Abstracts	Blog Posts	Social Media Posts
Air pollution	Accessibility	Administration	Career
Bicycles	Ageing	Automation	Food
Buildings	Antennas	Banking	Green spaces
Carbon emissions	Artificial Intelligence	Business	Railroad
Climate	Authentication	Crime	Road safety
Electric vehicles	Autonomous driving	Disaster management	Shopping
Healthcare	Big data	Education	Sidewalks
Housing	Blockchain	E-governance	Street art
Infrastructure	City architecture	Innovation	Suburban
IoT	Cyber-physical model	Investments	Urban planning
Noise pollution	Dashboard	Lighting	Urbanism
Parking space	Disaster management	Ports	
Public transportation	Edge computing	Railways	
Security	Education	RES	
Smart grid	E-governance	Road safety	
Sustainability	Fiber-optical	Sanitation	
Traffic management	Indicators	Surveillance	
Waste management	Industry	Taxes	
Water management	Innovation		
City	Intrusion detection		
	Logistics		
	Machine learning		
	Networks		
	Participation		
	Positioning		
	Privacy		
	Recommendations		
	Semantics		
	Sensors		
	Social media		
	Technology		
	Tourism		

Certain topics are core concepts within the smart city paradigm, such as “traffic management”, “IoT”, “air pollution”, “healthcare”, “security”, “smart grid”, etc., as evidenced by their presence across all datasets. Within the dataset of scientific publication abstracts, we have identified more specialized topics, including “edge computing”, “machine learning”, “semantics”, “antennas”, “blockchain”, and others, as detailed in the related list. It is interesting to focus on topics that are absent from the abstracts but prevalent in either the news blog posts or the social media posts datasets. For instance, topics like “business”, “investments”, and “banking” emerge in the news blog posts dataset, while topics related to the everyday lives of citizens, such as “sidewalks”, “green spaces”, “shopping”, “food”, “street art”, “suburban”, and “career” exclusively feature in the social media posts dataset. This phenomenon raises points for discussion, as elaborated upon in Section 4.

4. Discussion

In this study, our objective was to delve into the various aspects of the smart city concept. To achieve this, we performed a topic modeling analysis across three distinct datasets. Some topics identified through the analyses can be interpreted as overarching themes, while others can be seen as subthemes of those. For instance, “surveillance” could be regarded as a subtheme of “security”. Furthermore, there are topics that could be merged into a more comprehensive aspect. For instance, “smart grid”, “RES”, “lighting”, and even "electric vehicles" or "carbon emissions" could be integrated into a broader aspect called "energy management" Similarly, "electric vehicles", along with "traffic management", "parking space", "public transportation", "road safety", and "bicycles" could collectively constitute a larger theme named "mobility".

However, it is worth noting that not all themes express the same perspective. Upon examining the derived topics, we can discern three distinct perspectives within the smart city ecosystem: city applications, technology, and socio-economic considerations. For instance, "water management" represents a smart city application, "edge computing" pertains to a technological aspect, and "innovation" addresses a socio-economic perspective. Additionally, certain themes may hold significance across multiple perspectives. For instance, "security" is a theme that holds relevance within all perspectives.

4.1. Addressing RQ1

A number of themes and subthemes were extracted from the data using topic modeling and thematic analysis to address RQ1. Considering all the points mentioned above and harnessing the HC dendrograms alongside the topic heatmaps that we generated for the three datasets, we have discerned a range of themes, each tailored to a distinct perspective: city applications, technology, and socio-economic considerations. This information holds significant potential for informing smart city planning, administration, and service provision. In essence, the identified themes shed light on the specific aspects of smart cities that need to be taken into account during the design and management phases to effectively address various challenges. The findings are visualized on a thematic map in Figure 23.

From the perspective of city applications (highlighted in green in Figure 23), we have identified the following predominant aspects of the smart city:

Aspect 1: Mobility

As the population grows, a host of challenges arise, including traffic congestion, heightened emissions, and both environmental and economic concerns. In light of these issues, the implementation of pertinent strategies by smart cities becomes paramount in addressing the escalating problems tied to traffic. The mobility aspect encompasses themes related to the movement of people and goods within the city. It includes topics such as traffic management, electric vehicles, parking space, public transportation, road safety, bicycles, railroad, ports, positioning, autonomous driving, and sidewalks.

Aspect 2: Energy

Energy management undoubtedly stands as a pivotal factor for every city. Any instance of power failure is untenable when considering the construction of a smart city. Consequently, a smart city must undertake the following tasks: Firstly, it augments existing power systems via cutting-edge designs, streamlined service automation, supply monitoring and control, and the inclusion of charging and discharging facilities. Secondly, it promotes energy-conscious practices among users, imparting cost-effective energy usage strategies [55] and suggesting alternative options to facilitate informed decisions. Thirdly, it unifies all energy resources [56] within a singular framework. The energy aspect pertains to electricity and energy management within the smart city. It covers themes like smart grids, renewable energy sources, carbon emissions reduction, and advanced lighting solutions.

Aspect 3: Infrastructure

The smart city infrastructure serves as the initial stride toward establishing the comprehensive framework and architecture of a smart city. The incessantly growing need for resources, propelled by economic advancement and global population growth, necessitates a continuous commitment to the endeavors associated with sustainable development. Themes in this aspect encompass water management, buildings, housing, and suburban and city architecture.

Aspect 4: Environment

The environment plays a paramount role in the sustainable development of smart cities. A clean, pollution-free, and hazard-free environment is crucial for humanity's progress and the well-being of cities. Environment addresses the smart city's integration within a sustainable ecosystem. This aspect covers themes such as air pollution, green spaces, sustainability, waste management, noise pollution, disaster management, and climate.

Moving on to the technology perspective (highlighted in blue in Figure 23), we identified the following dominant aspects of the smart city:

Aspect 5: Internet of Things (IoT)

The internet has revolutionized our daily routines by optimizing resource utilization and enhancing our quality of life. The IoT, in turn, has transformed ordinary objects into intelligent devices capable of internet-based communication [9]. Smart city designs are replacing conventional urban planning approaches, incorporating wireless networks for smart city monitoring. Driven by AI-based methods, IoT devices and wireless sensor networks (WSNs) predominantly employ hard (or dedicated) sensing as the primary sensing paradigm in numerous smart city applications [57]. IoT encompasses themes related to the electronic infrastructure of the smart city. It includes topics like the cyber-physical model, technology, networks, sensors, edge computing, fiber-optical networks, and antennas.

Aspect 6: Data

In the smart city context, generating analytics can yield advanced insights, foster a deeper comprehension of urban phenomena, and facilitate the formulation of evidence-based urban strategies. The distributed setting utilized in handling big data for the smart city has the potential to present challenges in data storage and processing [58]. The exploration of meaningful patterns and correlations within city-service facilities using data mining (DM) methods has progressively gained prominence as a pivotal research domain. Data pertains to the technologies for gathering, managing, storing, processing, and analyzing the most valuable asset of our times: data. This aspect covers themes such as big data analytics, AI, ML, recommendation systems, dashboards, and semantics.

From a socio-economic perspective (colored in orange in Figure 23), we identified the following dominant aspects of the smart city:

Aspect 7: City planning and Administration

Governments and policymakers recognize the wealth of expertise and experience residing within the public and, therefore, encourage societal members to actively contribute to the decision-making process [59]. City planning and administration cover themes related to the design and management of a smart city. This aspect includes e-Governance, urban planning, indicators, urbanism, administration, taxes, social media, and participation.

Aspect 8: Business

Economists assert that a city can be deemed 'smart' when investments in communication infrastructure foster sustainable economic development, high quality of life, and prudent natural resource management while fulfilling the needs of residents, businesses, and institutions [60]. Business themes relate to economic activities within a smart city. Business includes innovation, logistics, industry, investments, banking, automation, and career.

Aspect 9: Security

Braun et al. [61] view security as a dynamic concept aimed at preventing harm to the smart city and its residents, encompassing both direct and indirect threats through digital and physical connections. Due to the widespread use of ICT and increasing interconnectivity, security themes play a critical role in safeguarding the smart city [62]. This aspect covers privacy, authentication, intrusion detection, surveillance, blockchain, and crime.

Aspect 10: People

Perhaps the most pivotal aspect of a smart city is its people. Quality of life encompasses the overall well-being of its residents, including education, healthcare, aging, accessibility, tourism, sanitation, shopping, food, and street art.

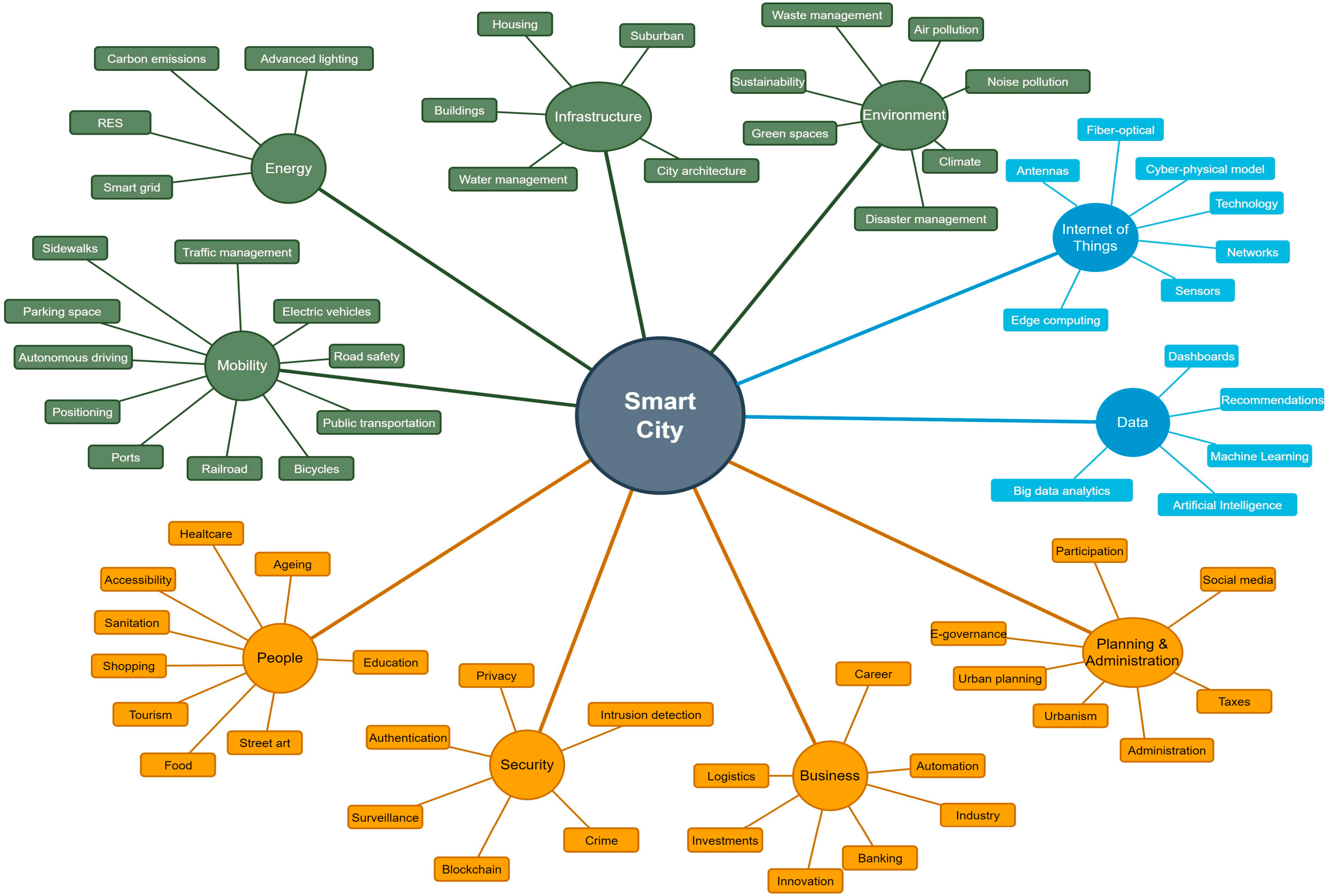

Figure 23. A thematic map of the smart city.

4.2. Addressing RQ2

The quality of life for citizens extends beyond the aspects we have explored thus far, encompassing various daily activities that have not received equivalent attention from the scientific community.

In the Reddit posts dataset, topics related to the everyday life of citizens, such as food, shopping, street art, sidewalks, and green spaces, have emerged. Interestingly, these topics received limited attention in the scientific publication dataset. This disparity highlights the influence of top-down approaches typically favored by experts in smart city design and underscores the necessity for a more citizen-centric approach.

The analysis of the news blog dataset emphasized the economic aspect of the smart city paradigm. It was somewhat anticipated that topics such as business, investments, banking, and jobs would predominate in this dataset, given its source from the online version of Business World magazine. However, what warrants attention is the complete absence of these themes in the other dataset, which could serve as a subject for further exploration and discussion.

4.3. Limitations

We acknowledge the limitations of this research. The analysis was conducted on relatively small datasets, particularly in the case of news blog posts and social media entries, where we examined only one source for each category: Business World News and Reddit, respectively. The internet contains a wealth of information, but there are various restrictions imposed by companies on access to this data. We selected these specific sources because they offered relatively unrestricted access and provided sufficient data for the analysis to produce reasonably reliable results. Clustering short texts in specific domains, like social media entries, is widely recognized as a challenging task. Li et al. [63] enumerate the reasons for this difficulty. First, the brevity of short texts results in low word frequency and sparse content, leading to unstable and inaccurate clustering outcomes. Second, within a narrow domain, there is often a significant overlap of less meaningful words, making it challenging to differentiate between sub-domains or create finely grained clusters. Additionally, the limited vocabulary size hinders the construction of a comprehensive word bag necessary for clustering algorithms to generate meaningful topic distributions.

In our analysis, we employed BERTopic, an unsupervised topic modeling library that produces unlabeled topics as output. As noted by Kumar et al. [64], the lack of labeled data can pose challenges in assessing the significance and relevance of unconventional topics, often making it difficult to distinguish them from noise or irrelevant topics. This challenge becomes especially pronounced when dealing with extensive and intricate datasets encompassing a wide array of topics. Furthermore, it is important to acknowledge the inherent subjectivity of qualitative analysis. While the collaboration between the two authors resulted in a comprehensive study, the reliability of the findings could be further enhanced with the input of additional experts in the field.

5. Conclusions

The concept of a smart city is about improving citizens' quality of life through advancements in information and communication technologies. A smart city is an interconnected system characterized by a high level of complexity. In this work, we investigated various aspects of the smart city paradigm to uncover its latent structure, yielding new prospects for the efficient administration of such a complex system. To accomplish this aim, we collected textual data from three distinct sources: scientific publication abstracts, news blog posts, and social media entries. For the analysis of this textual data, we devised a novel semi-automated methodology that incorporates topic modeling and thematic analysis. Our findings illustrate that the smart city domain is a complex system that must be examined and organized from three perspectives: applications, technology, and socio-economics. The analysis revealed ten themes/aspects of the smart city paradigm, including mobility, energy, infrastructure, environment, IoT, data, business, planning and administration, security, and people. When comparing the results from the three different datasets, we observed a lack of interest within the scientific community in certain aspects, such as business, as well as themes that are relevant to the everyday lives of citizens, including food, shopping, and green spaces.

This work contributes to the literature in three ways: (a) It enhances our comprehension of the structure of the smart city paradigm by examining its various aspects; (b) it reveals latent smart city topics that need more attention from the scientific community, opening new avenues for research; and (c) it proposes a semi-automated methodology for the thematic analysis of very large corpora by using the BERTopic topic modeling library. The insights presented in this work hold significant implications for scholars, practitioners, and public administrators engaged in smart city transformation efforts. Furthermore, this article sheds light on the latent aspects within the domain that require attention from stakeholders. Essentially, our work addresses the phenomenon known as the "two communities" or the "research-practice gap" [65], which pertains to how research can actively engage with and contribute to practical implementation. Additionally, this paper introduces a data-driven approach for conducting thematic analyses on extensive textual datasets, creating new opportunities for qualitative analysis within the vast sea of available data in our contemporary era.

Future Research

Our findings can serve as foundational material for future research endeavors focused on smart cities, offering a preliminary basis to support the development of the underlying ontology of a smart city. This, in turn, contributes to the construction of a comprehensive framework, such as a smart city knowledge graph. Researchers have the opportunity to explore novel avenues to expand upon this work, thereby addressing latent aspects of the smart city transformation and embracing a more citizen-centric perspective.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/fi16010003/s1, Interactive Visualization S1: Figure 5. Abstrats_intertopic_distance.html; Interactive Visualization S2: Figure 6. abstracts_dendrogram.html; Interactive Visualization S3: Figure 7. Abstracts_topics.html; Interactive Visualization S4: Figure 8. Abstracts_dendrogram_merged; Interactive Visualization S5: Figure 9. Abstracts_docviz.html; Interactive Visualization S6: Figure 10. Abstracts_heatmap.html; Interactive Visualization S7: Figure 11. News_intertopic.html; Interactive Visualization S8: Figure 12. News_dendrogram.html; Interactive Visualization S9: Figure 13. News_topics.html; Interactive Visualization S10: Figure 14. News_dendrogram_merged.html; Interactive Visualization S11: Figure 15. News_docviz.html; Interactive Visualization S12: Figure 16. News_heatmap.html; Interactive Visualization S13: Figure 17. Reddit_intertopic_distance.html; Interactive Visualization S14: Figure 18. Reddit_dendrogram.html; Interactive Visualization S15: Figure 19. Reddit_topics.html; Interactive Visualization S16: Figure 20. Reddit_dendrogram_merged.html; Interactive Visualization S17: Figure 21. Reddit_docviz.html; Interactive Visualization S18: Figure 22. Reddit_heatmap.html; Data S19: smart_city_papers.txt; Data S20: rssdata.txt; Data S21: redditdata.txt.

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

Funding: This research is co-financed by Greece and the European Union (European Social Fund-SF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014–2020" in the context of the project "Support for International Actions of the International Hellenic University", (MIS 5154651).

Data Availability Statement: The data presented in this study are available in Supplementary Materials.

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

Abbreviations

The following abbreviations are used in this manuscript.

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BW	Businessworld
c-TF-IDF	Class-based Term Frequency–Inverse Document Frequency
CTM	Contextualized Topic Models
HC	Hierarchical Clustering
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
ICT	Information and Communication Technologies
IoT	Internet of Things
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
ML	Machine Learning
NLP	Natural Language Processing
NMF	Non-negative Matrix Factorization
pLSI	Probabilistic Latent Semantic Indexing
SCC	Smart Cities Challenge
SVD	Singular Value Decomposition
TC	Topic Coherence
TD	Topic Diversity
TF-IDF	Term Frequency–Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection
URL	Uniform Resource Locator
WoS	Web of Science

Appendix A

Table A1. Topics of scientific publication abstracts.

Topic	Count	Representative Terms
−1	5377	city, smart, urban, data, cities, development, paper, information, service, system
0	418	iot, internet, things, application, service, smart, technology, architecture, platform, device
1	239	traffic, congestion, road, vehicle, control, intersection, flow, light, time, signal
2	235	waste, bin, garbage, collection, solid, management, recycling, disposal, trash, system
3	223	air, pollution, quality, monitoring, sensor, pollutant, pm25, concentration, environmental, health
4	212	blockchain, security, contract, decentralized, iot, transaction, blockchain-based, technology, privacy, ledger
5	199	big, data, analytics, processing, bda, smart, hadoop, analysis, architecture, framework
6	183	node, routing, network, protocol, wireless, sensor, wsn, packet, energy, algorithm
7	150	charging, electric, evs, ev, vehicle, station, battery, energy, vehicles, power
8	148	water, wastewater, supply, monitoring, management, groundwater, treatment, system, meter, distribution
9	147	parking, space, car, spot, driver, slot, system, lot, free, park
10	141	drone, uavs, uav, unmanned, aerial, flight, drones, vehicle, algorithm, flying
11	136	vehicle, vehicular, routing, communication, vanet, network, protocol, vanets, delay, scheme
12	130	lighting, street, light, lamp, led, streetlight, control, system, energy, leds
13	130	grid, power, energy, load, renewable, electricity, distribution, demand, microgrids, generation
14	127	evaluation, index, ranking, fuzzy, construction, criterion, weight, indicator, method, multicriteria
15	126	healthcare, health, patient, medical, patients, disease, care, diagnosis, elderly, monitoring
16	126	sentiment, twitter, tweet, social, medium, event, text, opinion, emotion, analysis
17	123	mobility, transport, sustainable, transportation, urban, smart, sustainability, study, city, public
18	123	bus, passenger, route, transportation, transport, travel, carpooling, public, problem, time
19	109	electricity, forecasting, energy, power, demand, load, consumption, building, machine, model
20	106	enterprise, architecture, city, smart, modeling, model, concept, system, integration, process
21	102	disaster, emergency, fire, response, evacuation, event, management, natural, relief, rescue
22	100	lorawan, lora, lpwan, network, low, wide, lowpower, sigfox, area, packet

Table A1. *Cont.*

Topic	Count	Representative Terms
23	99	design, de, game, space, playful, participatory, urban, ambient, la, theory
24	98	trust, security, vehicle, vehicular, attack, scheme, malicious, vanet, message, authentication
25	97	learning, deep, machine, ml, iot, data, dl, internet, things, neural
26	95	education, campus, university, learning, student, educational, teaching, skill, program, students
27	90	innovation, entrepreneurship, entrepreneurial, knowledge, collaborative, sector, open, literature, public, startup
28	88	energy, ped, district, building, peds, transition, decision, renewable, european, project
29	88	intrusion, attack, detection, ids, attacks, learning, security, deep, machine, malware
30	82	edge, computing, cloud, mec, offloading, resource, server, latency, application, mobile
31	80	fog, computing, cloud, latency, iot, paradigm, node, service, resource, edge
32	79	sustainable, sustainability, development, city, environmental, goal, goals, economic, ict, smart
33	76	effect, china, pilot, carbon, scp, construction, policy, green, emission, innovation
34	76	covid19, pandemic, health, crisis, epidemic, digital, virus, coronavirus, spread, disease
35	75	india, mission, indian, indias, rs, urbanization, urban, cities, 100, city
36	72	prediction, traffic, flow, neural, model, deep, congestion, temporal, graph, shortterm
37	72	security, cyber, cybersecurity, threat, vulnerability, information, attack, smart, dubai, city
38	72	image, detection, plate, object, pedestrian, license, reid, deep, reidentification, convolutional
39	70	noise, sound, acoustic, pollution, level, audio, measurement, microphone, monitoring, environmental
40	69	tourism, destination, brand, tourist, branding, destinations, hospitality, marketing, los, de
41	68	china, construction, chinas, chinese, development, city, smart, government, urbanization, smc
42	67	bicycle, bike, bikesharing, carsharing, sharing, cycling, transport, cyclist, ebikes, mobility
43	63	acceptance, intention, perceived, adoption, factor, trust, influence, equation, use, expectancy
44	61	localization, positioning, navigation, gps, indoor, ble, position, ekf, bluetooth, signal
45	61	mcs, crowdsensing, sensing, mobile, user, participant, recruitment, trustworthiness, participact, task
46	60	authentication, protocol, security, scheme, secure, user, device, proposed, mobile, iot
47	60	open, data, government, portal, public, city, transparency, initiative, new, datadriven
48	60	logistics, freight, delivery, lastmile, transport, mile, distribution, urban, cargo, last
49	60	discourse, urbanism, heritage, urban, songdo, narrative, metaverse, political, city, smart
50	55	video, surveillance, camera, edge, cctv, image, frame, processing, computing, retrieval
51	53	older, elderly, agefriendly, adult, care, frailty, ageing, senior, adults, health
52	50	ontology, semantic, data, web, linked, knowledge, domain, information, ontologies, personalised
53	50	citation, research, bibliometric, publication, analysis, topic, journal, literature, scientific, document
54	49	participation, citizen, citizens, engagement, involvement, eparticipation, public, government, citizenship, citizenry
55	47	urban, city, digital, development, concept, new, transformation, place, technology, social
56	46	resilience, resilient, disaster, seismic, hazard, earthquake, capacity, disasters, smartness, index
57	46	privacy, security, data, protection, citizens, issue, smart, trustless, information, risk
58	44	5g, vertical, network, technology, slice, gigabit, fifth, deployment, service, virtualization
59	44	climate, change, adaptation, roof, green, urban, planning, mitigation, european, earth
60	44	augmented, reality, ar, immersive, virtual, vr, experience, interaction, technology, 3d
61	44	3d, model, tree, shape, building, lidar, geological, citygml, virtual, scene
62	43	twin, digital, twins, dt, dts, physical, virtual, igb, replica, concept
63	42	indicator, assessment, dimension, smartness, city, smart, evaluation, smartnesssmart, salatiga, indicators
64	40	czech, republic, brno, concept, slovak, implementation, municipality, region, poland, slovakia
65	40	meter, energy, metering, power, grid, consumption, measurement, electrical, management, system
66	39	ai, artificial, intelligence, robot, robotics, xai, urban, human, research, development
67	38	face, recognition, facial, image, biometric, biometrics, feature, expression, fingerprint, emotion
68	37	mobility, human, mobile, train, irma, data, subway, transportation, cellular, trajectory
69	36	crime, police, criminal, prevention, safety, street, crimes, incident, prism, occurrence
70	36	manufacturing, 40, industry, production, product, industrial, instructions, instruction, revolution, factory
71	35	recommendation, recommender, user, poi, filtering, personalized, ecommerce, pois, preference, collaborative
72	35	accessibility, blind, disability, people, impaired, visuallyimpaired, wheelchair, accessible, urban, information
73	35	security, iot, threat, attack, vulnerability, iotbased, secure, risk, home, internet
74	32	governance, institutional, itg, actor, civil, organizational, perspective, smart, local, ecology
75	31	m2m, lte, simulation, iot, access, simulator, machinetomachine, communication, device, massive
76	31	antenna, ghz, band, frequency, mimo, gain, mmwave, db, dbi, polarization

Table A1. *Cont.*

Topic	Count	Representative Terms
77	30	indonesia, government, egovernment, governance, program, nusantara, ikn, regency, bandung, local
78	30	anomaly, detection, anomalous, outlier, series, tir, time, data, captchas, seismic
79	30	traffic, vehicle, road, detection, object, violation, sign, video, counting, image
80	28	risk, risks, assessment, credit, financial, project, social, management, construction, smart
81	27	trading, evs, energy, transaction, blockchain, p2p, charging, ev, sg, blockchainbased
82	27	optical, fso, turbulence, wavelength, transmission, gbps, channel, atmospheric, rofso, link
83	27	agent, multiagent, parcelrestbox, shopping, agents, customer, product, ordering, system, mots
84	27	building, bim, automation, buildings, bimiot, iot, integration, sensor, construction, floor
85	26	lidar, point, 3d, cloud, segmentation, object, synopsis, clouds, autonomous, vehicle
86	25	reinforcement, traffic, learning, signal, control, rl, agent, reward, intersection, deep
87	25	definition, concept, term, scp, understanding, literature, sc, city, smart, strategy
88	25	heat, energy, heating, renewable, thermal, solar, exergetic, cooling, district, pv
89	24	student, course, students, teaching, elearning, educational, learning, education, teacher, college
90	24	geospatial, information, gis, geographic, geoportals, cartographic, map, governance, spatial, thinkcities
91	22	energy, meter, consumption, home, saving, behaviour, ict, res, household, consumers
92	22	cyberphysical, cpss, cps, cyber, physical, systems, cpsc, world, system, cyberphysicalsocial
93	22	cultural, heritage, fourhospitality, relic, tourist, museum, intangible, mobile, preservation, tourists
94	22	capital, ic, intellectual, egyptian, policy, sids, managerial, study, case, framework
95	21	harvesting, selfpowered, piezoelectric, energy, nanogenerators, triboelectric, solar, power, mechanical, harvester
96	21	simulation, movement, agentbased, agent, abs, scenario, urban, model, fleet, replicate
97	20	dashboard, visualization, dashboards, visual, data, adaptable, visualize, urban, analyze, tvdp
98	20	pattern, trajectory, clustering, multidensity, mining, discover, spatiotemporal, correlated, hotspot, densitybased
99	20	federated, fl, learning, edge, privacy, training, server, centralized, fedtm, data
100	20	bim, gis, construction, ifc, shapefile, modeling, information, cim, integration, building
101	19	rainfall, flood, planting, hyperspectral, prediction, weather, image, crop, meteorological, arrangement
102	19	urbanism, digital, platform, labor, torontos, alphabet, politics, waterfront, political, sidewalk
103	19	business, value, model, service, canvas, bmes, framework, bm, case, smart
104	19	activity, recognition, har, human, home, activities, recognize, accuracy, sensor, labeling
105	18	blockchain, vehicle, vehicular, consensus, transaction, vehicles, iov, blockchainbased, security, secure
106	17	gamification, gamified, game, coremm, participation, citizen, engagement, codesign, public, behaviour
107	17	autonomous, vehicle, driving, fagvs, traffic, fagvinscf, venus, avs, scc, validation
108	17	parking, occupancy, prediction, space, availability, onstreet, free, min, slot, traffic
109	17	radio, aps, wifi, cell, channel, mobile, spectrum, converged, scheme, ddsa
110	16	health, food, ph, cancer, ugss, physical, healthy, social, environmental, sphec
111	16	governance, ethical, calzada, datadriven, data, policy, cooperatives, technostakeholders, paneuropean, barcelona
112	16	maturity, model, city, depok, assessment, south, level, colombia, smart, dimension
113	16	video, surveillance, detection, violence, clip, intrusion, neural, anomaly, vd, ivs
114	16	social, place, locationbased, people, clustering, tourist, density, activity, profiling, venue
115	15	lab, living, lls, labs, innovation, virtual, initiative, turin, torino, vending
116	15	egovernance, egovernment, government, india, governance, eservices, chapter, stuttgart, dubai, maturity
117	15	shm, structural, monitoring, sensor, structures, bridges, corrosion, wsan, health, ischm
118	15	crash, road, safety, collision, age, hard, driver, rearend, walkway, telematics
119	15	qol, csr, perception, citizens, scqol, scd, csgscs, quality, life, young
120	15	russian, russia, sphere, development, tyumen, tire, concept, tatarstan, petersburg, kazan
121	15	parking, camera, space, lot, occupancy, image, vacant, gate, detection, car
122	14	video, streaming, qoe, bandwidth, qos, transcoding, routing, bitrate, quality, adaptive
123	14	data, aqueducte, sharing, api, xml, dgt, apis, source, aggregation, service
124	14	energy, renewable, electricity, grid, rei, efficiency, production, realizing, sources, prosumers
125	14	image, encryption, chaotic, security, medical, scheme, sst, ehr, images, healthcare
126	13	fiber, optical, fiberoptic, sensor, fault, monitoring, bragg, fibreoptic, strain, multimode
127	13	nbs, sustainable, indicator, sustainability, assessment, aquaponics, sdg, greenness, citys, environmental
128	13	pavement, road, rehabilitation, mams, condition, ipavement, pmi, rct, civil, surface
129	12	sensing, vcs, driveby, coverage, bus, automotive, msigc, metric, spatiotemporal, scanner

Table A1. *Cont.*

Topic	Count	Representative Terms
130	12	population, floating, om, information, functional, administration, operator, city, data, kpi
131	12	disruptive, technology, photonic, living, technologies, cities, sci, internet, discus, web
132	12	differential, optimization, bso, evolutionary, deepso, swarm, storm, brain, globalbest, particle
133	12	crowdsourcing, duplicate, crowdsensing, campaign, msmc, report, pcm4de, rhetoric, collective, review
134	12	ant, colony, swarm, algorithm, mapping, pheromone, aa, firefly, hop, foraging
135	11	noma, decoding, spectrum, ced, interference, lt, transmission, aoi, softcoap, lhmn
136	11	housing, settlement, ekurhuleni, informal, audit, development, recreational, welllocated, wlli, subsidized
137	11	dubai, geoict, oil, dubais, ahmadi, kuwait, arabian, political, saudi, arab
138	11	mvaas, query, raw, view, data, processing, streams, stream, materialized, i2dlv
139	10	protection, regulation, gdpr, personal, legal, eu, sll, law, general, data
140	10	abecities, cloud, access, scheme, attributebased, obfuscation, encryption, authorized, authorization, data
141	10	crsg, iot, graph, siot, clustering, common, network, discovery, attributed, gn
142	10	home, automation, hns, appliance, temperature, icsce, android, connected, system, user
143	10	sec, decarbonisation, energy, transition, sles, projects, barrier, cities4zero, decarbonization, sustainability
144	10	community, participation, surveillance, researcher, citizen, survey, coh, imago, right, nict

Table A2. Topics of news blog posts.

Topic	Count	Representative Terms
−1	1072	city, india, smart, said, also, project, digital, technology, data, government
0	303	city, smart, cities, urban, data, mission, technology, development, project, infrastructure
1	131	solar, power, energy, mw, renewable, rooftop, wind, capacity, said, electricity
2	127	secretary, officer, ias, department, charge, additional, kumar, commissioner, posted, appointed
3	126	water, supply, river, ganga, rs, drinking, treatment, crore, sewage, urban
4	75	metro, rail, railway, railways, transport, train, project, corridor, crore, rs
5	61	waste, plastic, recycling, management, solid, ewaste, environment, packaging, india, tonne
6	59	india, skill, manufacturing, msmes, industry, product, training, innovation, technology, ai
7	56	digital, government, egovernance, kerala, service, state, governance, citizen, india, department
8	48	state, investment, industrial, mou, said, government, signed, business, policy, minister
9	48	highway, road, project, km, highways, rs, crore, nhai, expressway, construction
10	47	hospital, covid19, health, patient, medical, care, pandemic, healthcare, patients, doctor
11	46	security, cyber, cybersecurity, threat, data, attack, business, organization, breach, enterprise
12	44	estate, real, unit, housing, property, market, sale, home, developer, q1
13	44	delhi, bus, metro, taxi, transport, card, station, uber, cab, dmrc
14	41	iot, device, aeris, m2m, connected, data, internet, things, business, solution
15	41	led, bulb, eesl, energy, light, street, lighting, ujala, saving, programme
16	41	toilet, sanitation, swachh, cleanliness, bharat, waste, clean, city, toilets, urban
17	40	surveillance, security, video, city, smart, safety, camera, police, safe, technology
18	39	air, pollution, quality, delhi, burning, smog, level, delhis, teri, tower
19	36	climate, energy, change, emission, global, india, 2030, clean, action, said
20	34	payment, banking, paytm, bank, digital, cash, upi, payments, merchant, transaction
21	34	wipro, customer, cloud, global, digital, aws, wipros, sap, company, service
22	33	housing, house, affordable, urban, scheme, assistance, construction, minister, said, crore
23	33	grid, power, meter, smart, distribution, energy, transmission, discoms, metering, electricity
24	32	kuljeet, chahal, gridconnected, encourage, informed, ndmc, medium, member, civic, recently
25	32	chahal, kuljeet, gridconnected, ndmc, encourage, informed, medium, member, civic, recently
26	32	police, cctns, criminal, crime, station, policing, delhi, portal, complaint, said
27	31	startup, startups, woman, hub, ecosystem, entrepreneur, telangana, innovation, thub, program
28	30	disaster, management, flood, risk, said, rain, earthquake, reduction, district, plan
29	26	charging, electric, ev, evs, vehicle, transport, station, cesl, public, vehicles
30	26	ndmc, smart, delhi, civic, municipal, council, grievance, connaught, city, app
31	24	green, building, sustainable, griha, energy, bamboo, construction, buildings, igbc, design
32	24	urban, city, cities, urbanization, development, global, population, building, like, india
33	23	dda, delhi, development, said, land, duda, work, flat, government, agency
34	23	smart, cities, bw, businessworld, city, sapna, bhardwaj, yes, idea, jury

Table A2. *Cont.*

Topic	Count	Representative Terms
35	21	rpa, automation, ai, uipath, robot, technology, human, data, process, intelligence
36	17	school, education, teacher, child, learning, learner, mentor, classroom, teaching, student
37	17	gst, tax, bill, rate, sabha, wages, estate, regime, benefit, real
38	16	wifi, station, railway, google, railtel, facility, hotspot, railways, bsnl, stations
39	16	business, customer, parking, covid19, crisis, employee, organization, digital, need, time
40	15	road, safety, accident, death, crash, transport, fatality, traffic, bus, corporates
41	13	climate, energy, plan, action, change, consumption, city, planyc, greenhouse, gas
42	12	capital, amaravati, andhra, master, singapore, pradesh, ap, chandrababu, plan, japanese
43	12	ntt, business, smbs, customer, cloud, digitalization, india, technology, transformation, data
44	12	port, coastal, gadkari, ports, sagarmala, inland, said, rs, crore, waterway
45	12	island, andaman, nicobar, kashmir, bagh, prime, modi, connectivity, minister, jsw
46	11	dell, emc, vxrail, storage, data, vmware, technologies, workload, hci, cloud
47	11	uk, minister, fadnavis, visit, summit, modi, british, maharashtra, madhya, said
48	11	cloud, hybrid, citrix, business, organization, microsoft, enterprise, cloudbased, respondent, azure
49	11	aiib, bank, adb, asian, percent, china, billion, agreement, finance, infrastructure
50	10	aluva, urban, approval, naidu, development, meeting, respect, minister, gc, unhabitat
51	10	solar, railways, power, station, panel, railway, plant, mw, depot, installed

Table A3. Topics of social media posts.

Topic	Count	Representative Terms
−1	850	city, urban, street, design, new, building, community, transportation, downtown, architecture
0	292	smart, cities, city, iot, tech, platform, technology, data, open, digital
1	209	planning, planner, urban, job, degree, career, planners, design, anyone, designer
2	90	housing, affordable, crisis, affordability, apartment, house, rent, renters, apartments, lessons
3	60	transit, bus, public, transportation, shuttle, free, system, america, commuters, beep
4	51	design, project, urban, software, comfort, im, writing, code, inspiring, engineered
5	45	rail, highspeed, light, speed, stations, amtrak, high, cost, station, trains
6	42	future, city, cities, tokyo, spongy, corporate, connected, small, itself, effected
7	40	light, france, sign, pedestrian, streetlight, lights, lighting, range, french, traffic
8	39	zoning, ordinance, housing, policy, commission, minneapolis, singlefamily, based, reform, code
9	38	street, art, streets, public, london, legitimacy, hate, vandalism, stroad, artistic
10	38	walkable, heat, walkability, sidewalks, extreme, neighborhood, tropical, climates, uphill, cardependency
11	37	bike, bikes, ebikes, important, cycling, cyclist, electric, ebike, driving, bicycle
12	35	parking, space, lot, richmond, minimum, repeal, fee, yard, driver, replacing
13	33	car, carfree, dependency, american, carcentric, gas, evs, eliminating, dependent, lets
14	32	career, job, advice, grad, school, bimonthly, graduate, thread, education, academiccareer
15	30	urbanism, city, modernism, amazons, return, urbanists, beyond, perfect, live, opinion
16	27	lane, bike, picture, left, protected, emergency, right, lanes, overpass, turn
17	25	suburban, rural, town, village, tx, sprawl, tiny, suburbs, areas, growth
18	24	noise, air, pollution, quality, quiet, sound, monitoring, reduce, din, mostly
19	20	park, france, parks, central, mixed, use, old, station, repurposing, shoups
20	20	interview, advice, internship, tips, first, expect, interviewing, position, prepare, job
21	19	california, californias, earthquake, flaw, role, law, granny, woe, easing, surprising
22	19	tree, trees, forest, usc, climate, summer, third, researcher, extreme, resident
23	18	office, conversions, converting, building, empty, 11,000, offices, that, hard, says
24	18	book, urban, recommendations, planning, books, designplanning, slums, politics, shape, story
25	18	korea, south, seoul, apartment, built, class, complex, jugong, supplied, wirye
26	17	road, widening, roads, stroads, highway, suburb, shrink, conventional, se, webcast
27	17	malls, mall, shopping, thoughts, bring, used, may, walmart, residences, substitute
28	17	japan, paradise, snow, tokyo, japans, vietnams, anticar, became, concrete, carcentric
29	15	deaths, death, safety, road, hasnt, fatality, hoboken, traffic, years, pedestrian
30	14	gentrification, food, condo, tastes, trends, displacement, politics, lead, development, neighborhoods
31	14	jacobs, jane, book, death, life, review, enjoy, american, hope, great

Table A4. Topics of social media posts.

Topic	Count	Representative Terms
32	13	lawn, water, lawns, grass, court, plant, climate, droughtfriendly, norm, dwindles
33	13	saudi, arabia, riyadh, desert, worlds, mega, theme, billion, modern, largest
34	13	aicp, experience, exam, pas, week, education, dtlls, diploma, det, eit
35	13	congestion, pricing, toll, vehicle, wins, everybody, impacts, york, environmental, plan
36	12	land, tax, value, property, amendment, rising, 14th, ownership, flooded, depend
37	12	skyscrapers, skyscraper, construction, buildings, reuses, towering, emotionally, straddling, 101, 2700
38	11	traffic, cameras, speeding, crashes, limit, control, reduce, light, speed, anticipating
39	10	build, construction, america, created, europe, honestlycan, bubbles, enuf, leaked, mainland
40	10	berlin, bchle, freiburg, germany, 5 min, useful, canal, 8 lane, constructing, berlins

References

1. Sharma, C.; Batra, I.; Sharma, S.; Malik, A.; Sanwar Hosen, A.S.M.; Ra, I.-H. Predicting Trends and Research Patterns of Smart Cities: A Semi-Automatic Review Using Latent Dirichlet Allocation (LDA). *IEEE Access* **2022**, *10*, 121080–121095. [CrossRef]
2. United Nations. *World Urbanization Prospects 2018—Highlights*; Department of Economic and Social Affairs: New York, NY, USA, 2019.
3. Nicolas, C.; Kim, J.; Chi, S. Natural Language Processing-Based Characterization of Top-Down Communication in Smart Cities for Enhancing Citizen Alignment. *Sustain. Cities Soc.* **2021**, *66*, 102674. [CrossRef]
4. Zarindast, A.; Sharma, A.; Wood, J. Application of Text Mining in Smart Lighting Literature—An Analysis of Existing Literature and a Research Agenda. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100032. [CrossRef]
5. Wang, C.; Steinfield, E.; Maisel, J.; Kang, B. Is Your Smart Inclusive? Evaluating Proposals from the U.S. Department of Transportation's Smart City Challenge. *Sustain. Cities Soc.* **2021**, *74*, 103148. [CrossRef]
6. Stimmel, C.L. *Building Smart Cities: Analytics, ICT, and Design Thinking*; CRC Press: Boca Raton, FL, USA, 2016.
7. Townsend, A.M. *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*; W.W. Norton & Company: New York, NY, USA, 2013.
8. Kim, B.; Yoo, M.; Park, K.C.; Lee, K.R.; Kim, J.H. A Value of Civic Voices for Smart City: A Big Data Analysis of Civic Queries Posed by Seoul Citizens. *Cities* **2021**, *108*, 102941. [CrossRef]
9. Kousis, A.; Tjortjis, C. Data Mining Algorithms for Smart Cities: A Bibliometric Analysis. *Algorithms* **2021**, *14*, 242. [CrossRef]
10. Kar, A.K.; Dwivedi, Y. Theory Building with Data-Driven Research—Moving Away from the "What" towards the "Why". *Int. J. Inf. Manag.* **2020**, *54*, 102205. [CrossRef]
11. Gillies, M.; Murthy, D.; Brenton, H.; Olaniyan, R. Theme and Topic: How Qualitative Research and Topic Modeling Can Be Brought Together. *arXiv* **2022**, arXiv:2210.00707.
12. Kumar, S.; Kar, A.K.; Ilavarasan, V. Applications of Text Mining in Services Management: A Systematic Literature Review. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100008. [CrossRef]
13. Zheng, Z.; Sieber, R. *Interpreting the Smart City Through Topic Modeling*; Springer Link: Montreal, QC, Canada, 2023; Volume 1, pp. 29–46.
14. Suyanto, A.H.; Djatna, T.; Wijaya, S.H. Mapping and Predicting Research Trends in International Journal Publications Using Graph and Topic Modeling. *Indones. J. Electr. Eng. Comput. Sci.* **2023**, *30*, 1201–1213. [CrossRef]
15. Braun, V.; Clarke, V. Using Thematic Analysis in Psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. [CrossRef]
16. Isoaho, K.; Gritsenko, D.; Mäkelä, E. Topic Modeling and Text Analysis for Qualitative Policy Research. *Policy Stud. J.* **2021**, *49*, 300–325. [CrossRef]
17. Chang, J.; Boyd-Graber, J.; Gerrish, S.; Wang, C.; Blei, D. *Reading Tea Leaves: How Humans Interpret Topic Models*; Bengio, Y., Schuurmans, D., Lafferly, J., Williams, C., Culotta, A., Eds.; Curran Associates, Inc.: Vancouver, BC, Canada, 2009; Volume 22, pp. 288–296.
18. Greene, D.; O'Callaghan, D.; Cunningham, P. *How Many Topics? Stability Analysis for Topic Models*; Calders, T., Esposito, F., Hullermeier, E., Meo, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8724, pp. 498–513.
19. Asmussen, C.B.; Møller, C. Smart Literature Review: A Practical Topic Modeling Approach to Exploratory Literature Review. *J. Big Data* **2019**, *6*, 93. [CrossRef]
20. Park, K.C.; Lee, C.H. A Study on the Research Trends for Smart City Using Topic Modeling. *J. Internet Comput. Serv.* **2019**, *20*, 119–128.
21. Wang, J.; Wang, M.; Song, Y. *A Study on Smart City Research Activity Using Bibliometric and Natural Language Processing Methods*; ACM: Guangzhou, China, 2021; pp. 1–7.
22. Lee, W.S. Analysing the Evolution of Interdisciplinary Areas: Case of Smart Cities. *J. Glob. Inf. Manag.* **2022**, *30*, 1–23. [CrossRef]
23. Esposito, G.; Terlizzi, A.; Guarino, M.; Crutzen, N. Interpreting Digital Governance at the Municipal Level: Evidence from Smart City Projects in Belgium. *Int. Rev. Adm. Sci.* **2023**, 1–17. [CrossRef]
24. Zheng, Z.; Sieber, R. Putting Humans Back in the Loop of Machine Learning in Canadian Smart Cities. *Trans. GIS* **2021**, *26*, 8–24. [CrossRef]

25. Alswedani, S.; Katib, L.; Abozinadah, E.; Mehmood, R. Discovering Urban Governance Parameters for Online Learning in Saudi Arabia During COVID-19 Using Topic Modeling of Twitter Data. *Front. Sustain. Cities* **2022**, *4*, 751681. [CrossRef]
26. Vargas-Calderón, V.; Camargo, J. Characterization of Citizens Using Word2vec Latent Topic Analysis in a Large Ser of Tweets. *Cities* **2019**, *92*, 187–196. [CrossRef]
27. Sinha, M.; Guha, S.; Varma, P.; Mukherjee, T.; Mannarswamy, S. *My City, My Voice: Listening to the Citizen Views from Web Sources*; ACM: Kolkata, India, 2019; pp. 52–60.
28. Saheb, T.; Dehghani, M.; Saheb, T. Artificial Intelligence for Sustainable Energy: A Contextual Topic Modeling and Content Analysis. *Sustain. Comput. Inform. Syst.* **2022**, *35*, 100699. [CrossRef]
29. Valença, G.; Moura, F.; de Sá, A.M. How Can We Develop Road Space Allocation Solutions for Smart Cities Using Emerging Information Technologies? A Review Using Text Mining. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100150. [CrossRef]
30. Srinivasa-Desikan, B. *Natural Language Processing and Computational Linguistics*; Packt Publishing Ltd.: Birmingham, UK, 2018.
31. De Oliveira Capela, F.; Ramirez-Marquez, E. Detecting Urban Identity Perception via Newspaper Topic Modeling. *Cities* **2019**, *93*, 72–83. [CrossRef]
32. Koukaras, P.; Tjortjis, C. Social Media Analytics, Types and Methodology. In *Machine Learning Paradigms. Learning and Analytics in Intelligent Systems*; Tsihrintzis, G., Virvou, M., Sakkopoulos, E., Jain, L., Eds.; Springer: Cham, Switzerland, 2019; pp. 401–427.
33. Rousidis, D.; Koukaras, P.; Tjortjis, C. Social Media Prediction: A Literature Review. *Multimed. Tools Appl.* **2020**, *79*, 6279–6311. [CrossRef]
34. Jeong, B.; Yoon, J.; Lee, J.-M. Social Media Mining for Product Planning: A Product Opportunity Mining Approach Based on Topic Modeling and Sentiment Analysis. *Int. J. Inf. Manag.* **2019**, *48*, 280–290. [CrossRef]
35. Egger, R. Topic Modelling: Modelling Hidden Semantic Structures in Textual Data. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*; Springer Nature: Cham, Switzerland, 2022; pp. 375–403.
36. Storopoli, J.E. Topic Modeling: How and Why to Use in Management Research. *Iberoam. J. Strateg. Manag.-IJSM* **2019**, *18*, 316–338. [CrossRef]
37. Nikolenko, S.; Koltcov, S.; Koltsova, O. Topic Modeling for Qualitative Studies. *J. Inf. Sci.* **2017**, *43*, 88–102. [CrossRef]
38. Mohr, J.; Bogdanov, P. Topic Models: What They Are and Why They Matter. *Poetics* **2013**, *41*, 545–569. [CrossRef]
39. Ogunleye, B.; Maswera, T.; Hirsch, L.; Gaudoin, J.; Brunsdon, T. Comparison of Topic Modelling Approach in the Banking Context. *Appl. Sci.* **2023**, *13*, 797. [CrossRef]
40. Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T.; Harshman, R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
41. Hoffman, T. *Probabilistic Latent Semantic Analysis*; Université de Montréal: Stockholm, Sweden, 1999.
42. Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
43. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv* **2022**, arXiv:2203.05794.
44. Wang, Z.; Chen, J.; Chen, J.; Chen, H. Identifying Interdisciplinary Topics and Their Evolution Based on BERTopic. In *Scientometrics*; Springer: Berlin/Heidelberg, Germany, 2023.
45. Mazzei, D.; Ramjattan, R. Machine Learning for Industry 4.0: A Systematic Review Using Deep Learning-Based Topic Modelling. *Sensors* **2022**, *22*, 8641. [CrossRef]
46. Thakur, O.; Saritha, S.K.; Jain, S. *Topic Modeling, Sentiment Analysis and Text Summarization for Analyzing News Headlines and Articles*; Khare, N., Tomar, D.S., Ahirwal, M.K., Semwal, V.B., Soni, V., Eds.; Springer: Cham, Switzerland, 2022; pp. 220–239.
47. Reimers, N.; Gurevych, I. *Sentence-Bert: Sentence Embeddings Using Siamese BERT-Networks*; Association for Computational Linguistics: Hong Kong, China, 2019; Volume 1, pp. 3982–3992.
48. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
49. McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861. [CrossRef]
50. Campello, R.J.G.B.; Moulavi, D.; Sander, J. *Density-Based Clustering Based on Hierarchical Density Estimates*; Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7819, pp. 160–172.
51. Cai, G.; Sun, F.; Sha, Y. *Interactive Visualization for Topic Model Curation*; ESIDA: Tokyo, Japan, 2018; Volume 2068.
52. McHugh, M.L. Interrater Reliability: The Kappa Statistic. *Biochem. Medica* **2012**, *22*, 276–282. [CrossRef]
53. Kherwa, P.; Bansal, P. Topic Modeling: A Comprehensive Review. *EAI Endorsed Trans. Scalable Inf. Syst.* **2019**, *7*, 1–16. [CrossRef]
54. Guest, G.; MacQueen, K.; Namey, E. *Applied Thematic Analysis*; Sage Publications: Thousand Oaks, CA, USA, 2012.
55. Khatavkar, N.; Naik, A.A.; Kadam, B. Energy Efficient Street Light Controller for Smart Cities. In Proceedings of the 2017 International Conference on Microelectronic Devices, Circuits and Systems (ICMDCS), Vellore, India, 10–12 August 2017; pp. 1–6.
56. Hoang, A.T.; Pham, V.V.; Nguyen, X. Integrating Renewable Sources into Energy System for Smart City as a Sagacious Strategy towards Clean and Sustainable Process. *J. Clean. Prod.* **2021**, *305*, 127161. [CrossRef]
57. Liu, Y.; Yang, C.; Jiang, L.; Xie, S.; Zhang, Y. Intelligent Edge Computing for IoT-Based Energy Management in Smart Cities. *IEEE Netw.* **2019**, *33*, 111–117. [CrossRef]
58. Singh, P.; Nayyar, A.; Kaur, A.; Ghosh, U. Blockchain and Fog Based Architecture for Internet of Everything in Smart Cities. *Future Internet* **2020**, *12*, 61. [CrossRef]

59. Siyam, N.; Alqaryouti, O.; Abdallah, S. Mining Government Tweets to Identify and Predict Citizens Engagement. *Technol. Soc.* **2020**, *60*, 101211. [CrossRef]
60. Makarchenko, M.; Nerkararian, S.; Shmeleva, I. How Traditional Banks Should Work in Smart City. In *Digital Transformation and Global Society. DTGS 2016. Communications in Computer and Information Science*; Chugunov, A., Bolgov, R., Kabanov, Y., Kampis, G., Wimmer, M., Eds.; Springer: Cham, Switzerland, 2016; Volume 674, pp. 123–134.
61. Braun, T.; Fung, B.; Iqbal, F.; Shah, B. Security and Privacy Challenges in Smart Cities. *Sustain. Cities Soc.* **2018**, *39*, 499–507. [CrossRef]
62. Almeida, F. Prospects of Cybersecurity in Smart Cities. *Future Internet* **2023**, *15*, 285. [CrossRef]
63. Li, C.; Lu, Y.; Wu, J.; Zhang, Y.; Xia, Z.; Wang, T.; Yu, D.; Chen, X.; Liu, P.; Guo, J. LDA Meets Word2Vec: A Novel Model for Academic Abstract Clustering. In Proceedings of the Companion Proceedings of the Web Conference, Lyon, France, 23–27 April 2018; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2018; pp. 1699–1706.
64. Kumar, M.; Rani, R.; Botarelli, M.; Epiophaniou, G.; Maple, C. Science and Technology Ontology: A Taxonomy of Merging Topics. *arXiv* **2023**, arXiv:2305.04055.
65. Walker, R.; Zhang, J.; Chandra, Y.; van Witteloosyuijn, A. Topic Modeling the Research-Practice Gap in Public Administration. *Public Adm. Rev.* **2019**, *79*, 931–937. [CrossRef]

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