The role of Big Data in Climate research

Andreea-Mihaela NICULAE[1], Alin-Gabriel VĂDUVA[1]
[1]Department of Economic Informatics and Cybernetics
andreea.niculae@csie.ase.ro, alin.vaduva@csie.ase.ro

This study explores the growing impact of Big Data in climate change research through a novel approach that combines Big Data analytics with text mining, natural language processing (NLP), and Latent Dirichlet Allocation (LDA). We analysed 7,145 open-access publications from 2011 to 2022 sourced from the Web of Science. Our work highlighted key themes such as urban health, smart technologies, and algorithmic modelling. We observed substantial growth in the use of Big Data in climate research up until 2022, followed by a surprising decline in 2023 that calls for further investigation. Sentiment analysis of the abstracts showed mostly neutral tones, although some exceptions revealed diverse perspectives. This research offers valuable insights into current trends, demonstrating the strength of an integrated analytical approach and the evolving role of Big Data in climate change research. The unexpected downturn in 2023 suggests a shift in research priorities, warranting further exploration.
Keywords: *Big Data, Climate-Change, Bibliometrics, Latent Dirichlet Allocation*

1 Introduction

Climate change is a complex, urgent challenge that calls for new ways to understand and mitigate its effects. Given the vast amounts and variety of climate-related data available today, advanced analytical techniques are essential for drawing useful insights from this information. This study examines the expanding role of Big Data in climate change research, analysing its applications and potential to deepen our understanding of this global issue. Using a dataset of 7,145 open-access publications from the Web of Science (2011-2022), we applied a range of analytical methods, Big Data processing, text mining, natural language processing (NLP), and Latent Dirichlet Allocation (LDA), to identify major research themes. Our analysis points to a substantial increase in climate-focused Big Data research, with clear clusters around urban health, the integration of smart technologies, and advanced modelling techniques. A decrease in publications from 2023, however, raises questions for further investigation. While the application of Big Data in climate change research is rapidly expanding, a comprehensive analysis of its impact and the associated research trends remains limited. This study directly addresses this gap by examining a large dataset of publications, uncovering key thematic areas, and analysing evolving research sentiments. Our findings offer insights into the current state of the field, highlighting areas of strength and weakness, thus contributing to a more informed and strategic approach to climate change research.

The paper is structured as follows: it begins with an introduction, followed by a literature review, a detailed description of the methodology and data sources, and concludes with the results and final insights.

2 Literature Review

The intersection of climate change and Big Data has become a rapidly expanding research area, driven by advancements in data processing and analytics [1]. Big Data's ability to manage vast amounts of diverse, real-time information allows researchers to model complex climate systems [2], predict extreme weather events, and even develop sustainable solutions.

The authors of [1] categorized Big Data applications in climate change research into five key areas through a selective review of

over 100 studies in 2019: sustainable urban planning, natural disaster assessment, smart farming, energy efficiency, and other advanced supports. For example, Big Data enables predictive modelling for natural disasters, supports precision agriculture to optimize yields and reduce environmental impacts, and enhances energy efficiency through real-time monitoring. These applications demonstrate Big Data's potential to bolster sustainable practices and increase resilience against climate-related challenges.

Many studies also explore the role of machine learning (ML) in climate change research [3]. While ML and Big Data are distinct domains, they often intersect in climate studies, with ML techniques frequently applied to analyze extensive datasets. The authors of [3] present a bibliometric analysis on ML applications in climate research, examining trends in publication growth, journal prominence, citation patterns, and collaboration networks. Their findings reveal that ML is widely used across various climate topics, including agriculture (e.g., crop predictions), natural disaster modeling (e.g., floods), and weather forecasting (e.g., rainfall predictions).

In addition to bibliometric analyses, the intersection of big data and climate change can be explored through topic modeling, specifically using Latent Dirichlet Allocation (LDA). LDA can analyze various text sources, such as bibliometric abstracts, news articles, social media posts, and other large text collections, to uncover underlying topics and trends. LDA is a useful method in climate change studies as it can capture public sentiment, identify diverse perspectives, and even help categorize broad topics. One research [4] uses LDA on big data obtained from youtube news videos of climate change-related domains in South Korea, where the authors discovered, among others, that many people are scared of climate change, some are relatively aware of environmental issues, few people make serious efforts to suggest alternatives, and that people think about climate change in many ways: social, politics, international, unrealistic, global, lifestyle-centric, and escapist.

3 Methodology

To analyse the large and complex body of research data on climate change and big data, we applied several complementary techniques, including Big Data analytics, Text Mining, Natural Language Processing (NLP), and Latent Dirichlet Allocation (LDA). Together, these methods help us process, categorize, and identify patterns within extensive text datasets, shedding light on recurring research trends and key themes. The following sections describe each part of this approach and its role in analysing the data.

3.1 Big Data

Big Data is a popular term [5] with an increase of usage in recent year due to advancements in computing power that allow for handling vast datasets. This term encompasses a multitude of concepts [6], from the well-known **4 V's**, Volume, Velocity, Variety, and Veracity, to digital techniques and large-scale data that traditional databases cannot accommodate. Big Data generally refers to large volumes of mostly unstructured data from various sources, such as individuals, machines, or sensors. This unstructured data can take many forms, including numerical values, large bodies of text, images, videos, geospatial information, and other types of rapidly generated data.

3.2 Text Mining

To fully leverage the high volume of Big Data, various data analysis techniques have been developed over the years. One notable technique is related to data mining, specifically text mining, which focuses on extracting information and identifying patterns from large volumes of text documents [7]. Text mining serves several

purposes, including document classification and clustering, information retrieval and extraction, web mining, concept extraction, and natural language processing (NLP). These techniques capture the essence of textual data and transform information into valuable knowledge.

3.3 Natural Language Processing

Natural Language Processing (NLP) is a field that combines elements of artificial intelligence and linguistics, using machine learning to interpret and generate human language [8]. NLP employs both supervised and unsupervised learning algorithms to handle a range of tasks, including speech recognition, text analysis, and sentiment analysis. By analyzing language structure (using tokenization, parsing, and syntactic analysis) and semantic relationships between words, NLP models can detect patterns, categorize text, and identify themes within large bodies of text. The field also applies linguistic principles to capture nuances in language, such as context, syntax, and sentiment, allowing for applications like language translation, entity recognition, and topic modeling.

3.4 Latent Dirichlet Allocation

A popular method from the NLP family is Latent Dirichlet Allocation (LDA), an unsupervised generative probabilistic model used to discover hidden topics within a significant body of text (or corpus) [9]. In this study, LDA helps identify underlying themes in climate change and big data research by grouping related words into topics. LDA operates by assigning probabilities to words for each topic, under the assumption that each document contains multiple topics. Key preprocessing steps, such as tokenization, removing stop words, and lemmatization, prepared the data for LDA. Model parameters, including the number of topics, were selected to optimize interpretability. A coherence score metric, which measures the semantic similarity among words in each topic, was used to assess topic quality and refine model parameters. Higher coherence scores indicate more meaningful and reliable topics, making the model's results useful for further analysis.

4 Data source

Our research utilizes data retrieved from the Web of Science (WoS) Core Collection database, accessed through an institutional account.

Table 1. Web of Science Core Collection Data Extraction Query

Web of Science filters	**Query**
All Fields	ALL=(clim* chang*) OR ALL=(glob* warm*) OR ALL=(glob* heat*) OR ALL=(clim* warm*) OR ALL=(greenhouse) OR ALL=(environm*) OR ALL=(clim*)
Topic	TS=("big data")
Publication Type	DT==("ARTICLE")
Open Access	OA==("OPEN ACCESS")
Language	LA==("ENGLISH")
Status	NOT (EN==("RETRACTED PUBLICATION"))
Publication Year	NOT(PY==("2025") OR PY==("2024"))

Using the complex query outlined in Table 1, we obtained bibliographic information for 7,145 publications. Publications from 2024 and 2025 were excluded as they are still in the publication process. We focused exclusively on Open Access articles to ensure easy and free access to the full content for any further assessments. The principal topic of this research is Big Data. To create a comprehensive view of climate change research we selected a variety of terms covering different aspects of the topic.

These include "climate change," "global warming," "global heating," "climatic warming," "greenhouse," "environment," and other derivatives of the terms.

5 Results and interpretation

Initial screening of the dataset revealed 7,145 records and 72 variables. However, 27 variables contained only missing values and were eliminated from the analysis. Among the remaining columns, 17 had more than 50% missing values, making them less useful for further analysis, and these were also removed. The rest of the columns describe various bibliographic aspects, with key variables including Publication Type, Authors, Article Title, Author Keywords and Keywords Plus, Abstract, Affiliations, number of Citations, Web of Science Categories, and Research Areas. For all the key variables, records with missing data were eliminated, ensuring only original information is being analyzed.

Upon further inspection, we found that none of the 7,145 publications were classified as Highly Cited or Hot Paper. This suggests that while the obtained publications contribute to the field of big data in climate change research, they have not achieved a considerable impact or recognition in any domains.

To better understand the obtained dataset, Figure 1 depicts the rising number of publications according to the top 5 most popular research fields found in big data and climate research. Originally, the data spans from 2011 to 2023, with 2011 having only one publication, unfortunately not found in the top 5 research fields. While the query was set to all years except 2024 and 2025, it is important to note that the oldest open-access article focused on the use of big data in climate research is recent, 2011 signaling a rather new research focus. The publication count follows a positive trend, resembling exponential growth, and peaks in 2022, followed by a significant drop in 2023. This decrease may reflect a shift in research interest toward other topics, or perhaps researchers have begun using different terminology to address similar issues. Further analysis is needed to better understand the decline observed in the most recent year.

The dataset contains numerous research areas, with Figure 1 presenting only the top 5 most popular of them. The overall increase in publications, across all research areas, suggests increasing research activity in them and heightened interest to address climate-related issues through diverse research approaches making use of the advantages of using big data. The most popular area is *Computer Science*, followed closely by *Environmental Sciences & Ecology*, signifying their relevance in climate research. *Engineering* and *Science & Technology – Other Topics* show substantial increases along years, but not as notable as the two topics mentioned before. *Telecommunications* shows the lowest number of publications relative to the other fields, but still demonstrates growth.

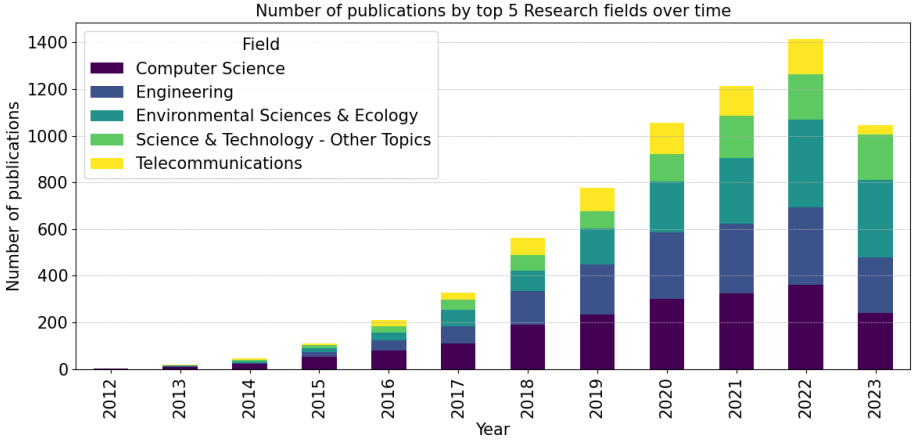

Fig. 1. Number of big data and climate research publications by top 5 research fields

Having examining the growth and distribution of publications over time, we now turn our attention to the content of these publications through text mining techniques, such as generating Word Clouds using Keywords Plus, performing sentiment analysis, or employing Latent Dirichlet Allocation. These techniques allow us to obtain a more in-depth understanding of key themes and patterns found in the big data for climate research publications.

The first analysis we perform is on Keywords Plus, a value created by the WoS database by exploring the articles and extracting common key areas. These values have the same writing style and are easier to use than Author Keywords, which, due to the human's way writing, might include derivations or combinations of terms, making them harder to combine and interpret. Figure 2 presents a Word Cloud obtained from the Keywords Plus of the 7,145 publications. *Big Data* is the dominant theme, seen from its centrality, followed by *Model* and *Management*, indicating a focus on developing *Frameworks* for managing the high volume of climate data effectively. Keywords such as *Prediction*, *Classification*, *Algorithm*, *Performance*, and *Optimization* suggest common analytical techniques used in the field reflecting the multiple methodological approaches being employed in this research. Terms like *Challenges* and *Innovation* suggest that big data presents opportunities in this domain, but it also brings complexities that need to be addressed in climate research. The inclusion of terms such as *Internet, Future, City, Environment* and *Health*, indicates the broader context in which big data and climate change are discussed, linking them to societal impacts. Overall, this word cloud illustrates the variety of elements surrounding big data in the context of climate change, highlighting the role of advanced analytical techniques to uncover insights and address challenges in this critical area of study.

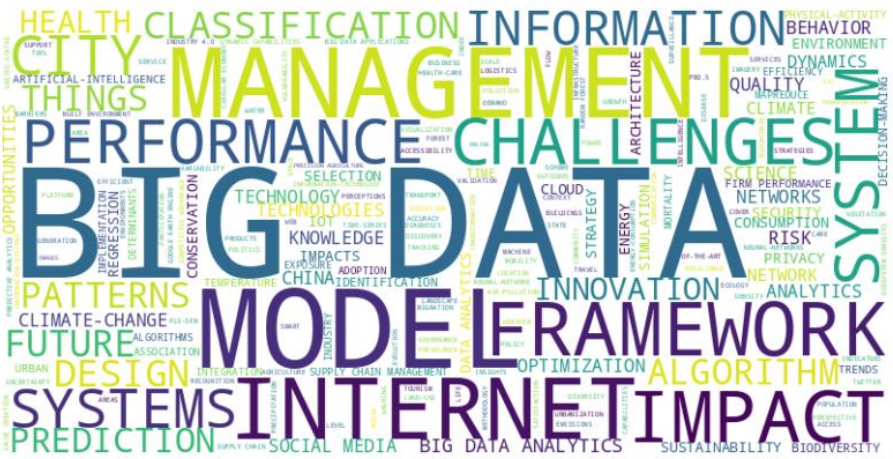

Fig. 2. Word Cloud from Keywords Plus

Using the Natural Language Toolkit (NLTK) package known as SentiWordNet, sentiment analysis was performed on the abstracts, yielding scores for positive and negative sentiments within the texts. In Figure 3, we observe the distribution of both positive and negative scores for each article. Most points are clustered in the lower ranges, suggesting that many abstracts have neutral or low sentiment overall. There appears to be a slight trend where higher positive sentiment scores correlate with increasing negative sentiment scores; however, many abstracts with high positive scores maintain low negative scores. The plot also reveals several outliers, indicating abstracts with exceptionally high sentiment scores. These publications should be further investigated for a better understanding of the

implications of big data in climate research. Overall, the results of SentiWordNet analysis indicates that while many abstracts are relatively neutral or slightly positive, there are unique cases that stand out, reflecting varying sentiments in climate change-related discussions.

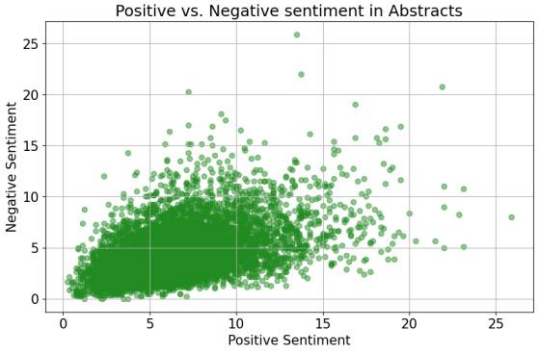

Fig. 3. Analysis of Positive and Negative sentiments from abstract mining

The last text mining approach we use is employing Latent Dirichlet Allocation on all the article's Abstracts to discover common themes. Before showing the results, it is important to mention that LDA's coherence scores for different number of topics were all below 50%, as seen in Table 2. For ease of readability and understanding the hidden topics, we have chosen to firstly represent 3 topics. By increasing the number of topics, while the coherence score increases, the visual representation of topics starts overlapping more and more. With 3 topics, there is no overlap in the plot. We will also present 7 topics to learn more about the implications of big data in climate research, while also showcasing the overlapping themes.

Table 2. LDA coherence scores for different number of topics

Number of topics	Coherence score (%)
3	40
4	39
5	41
6	43
7	48
8	47
9	47

The 3 chosen topics are named after the most important themes identified in them, as it can be seen in Figure 4. Topic 1, *Urban health analysis*, reflects the focus on urban areas and health factors, highlighting spatial analysis and results from studies concentrated on city environments. Topic 2, *Smart technology and Digital research*, emphasises the presence of research and technology, particularly in the context of digital systems and smart solutions, highlighting the advancement and application of modern technologies. Topic 3, *Algorithmic modelling and System development*, focuses on modelling methods, algorithms, and systems in the context of machine learning and research development. These word clouds capture the main themes of each topic, illustrating the varied yet interconnected aspects of climate change research and the use of big data. The prominent terms showcase the multidisciplinary nature of the field, revealing how areas like urban health, digital innovation, and algorithmic methods combine to tackle today's complex environmental challenges.

Fig. 4. LDA Word Clouds for 3 topics

Figure 5 presents LDA's Intertopic Distance Map for the three topics. The significant distance between the topics suggests that they address distinct areas within the larger research context. This separation indicates that each topic focuses on different themes or aspects, highlighting the diversity of research discourse surrounding climate change and big data.

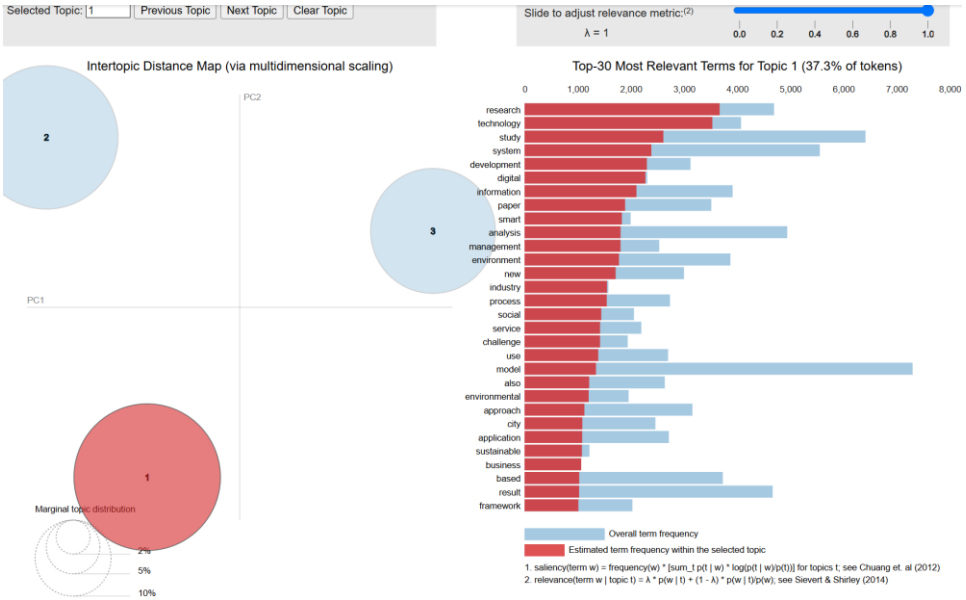

Fig. 5. LDA Intertopic Distance Map for 3 topics

Lastly, we present the obtained 7 topics, the number of topics with the highest coherence score. The new topics are similar to the ones previously obtained, while also containing new information, as seen in Figure 6: Topic 1, Smart technology and Digital management; Topic 2, Water resource management and Climate impact; Topic 3, Algorithmic Systems and Environmental modeling; Topic 4, Urban dynamics and Spatial activity patters; Topic 5, Social development and Digital research; Topic 6, Health prediction models and Risk analysis; Topic 7, Large dataset analysis and Learning methods.

These topics illustrate an interdisciplinary approach, with intersections between technology, social aspects, health, and environmental studies. For example, *Smart technology and Digital management* combines technological advances with management challenges, while *Social development and Digital research* considers societal impacts of technology. Topics focusing on specific application areas, such as *Water resource management and Climate impact* and *Health prediction models and Risk analysis*, suggest that researchers are increasingly applying big data to practical and critical issues, emphasizing the relevance of research in real-world contexts.

Figure 7 shows the Intertopic Distance Map for all seven topics, highlighting how they relate to one another within the research landscape. The topics in the lower half of the plot represent distinct areas of study with minimal thematic overlap, while those in the upper half share common themes, suggesting overlapping concepts or methodologies. This visualization offers useful insights into the connections between various areas of climate change and big data research.

The 7 topics outline a broad and varied landscape of research connecting climate change and big data. They show how different methods and application areas come together to deepen our understanding of complex environmental issues. This creates a useful framework for tackling

climate change through interdisciplinary research and innovative data-driven approaches, in contrast to the narrower themes seen in the initial analysis.

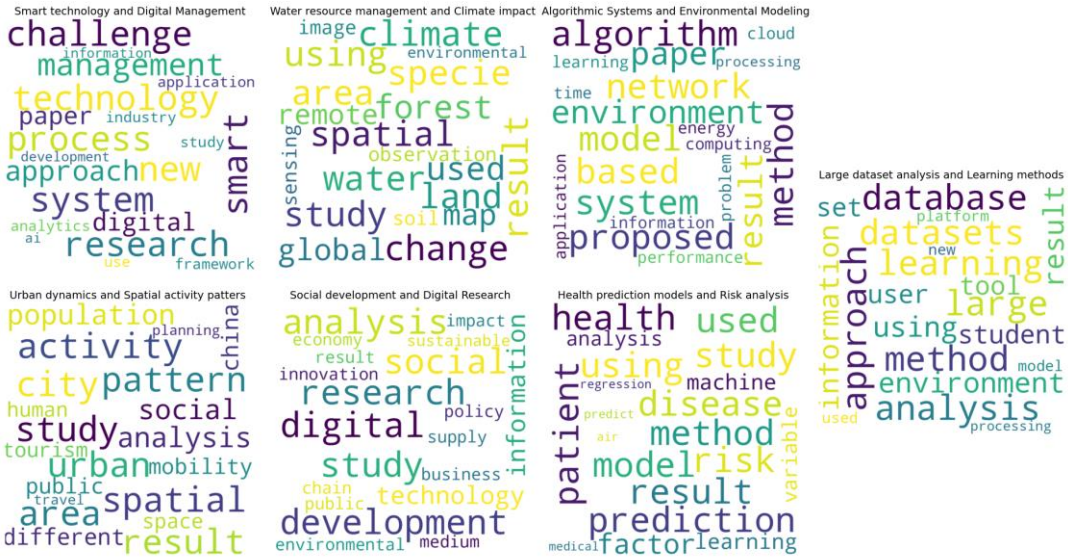

Fig. 6. LDA Word Clouds for 7 topics

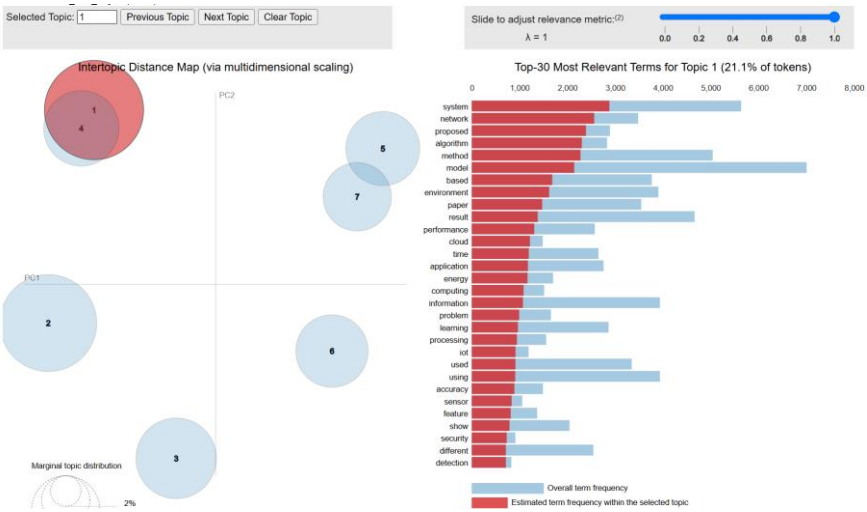

Fig. 7. LDA Intertopic Distance Map for 7 topics

6 Conclusions

This research examined the growing role of Big Data in climate change studies, using a methodology that combined Big Data analytics, text mining, natural language processing (NLP), and Latent Dirichlet Allocation (LDA). We analyzed 7,145 open-access publications from the Web of Science, covering the years 2011 to 2022 (excluding 2024 and 2025 due to ongoing publication). Our findings showed a clear increase in relevant publications from 2011 to 2022, followed by a decline in 2023. This drop raises questions about whether it reflects a shift in research priorities or a data anomaly, and further study is needed to understand this change.

In examining keyword frequencies, 'Big Data,' 'Model,' and 'Management' stood out as dominant themes, pointing to a strong emphasis on building effective frameworks for managing and analyzing large climate datasets. Sentiment analysis of the abstracts revealed mostly neutral sentiment, though with a few notable exceptions suggesting diverse viewpoints on Big Data's role in the field. LDA topic modeling, despite lower coherence scores,

helped uncover key thematic clusters, including urban health analysis, the integration of smart technologies, and advanced algorithmic models.
Overall, these findings highlight the essential role Big Data plays in addressing climate change challenges and underscore the need for interdisciplinary collaboration across scientific domains. Future work should focus on closely investigating the 2023 publication dip, refining sentiment analysis by including full-text data, enhancing the LDA model through parameter optimization, and addressing potential biases due to missing data.

7 Acknowledgment

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI - UEFISCDI, project number COFUND-CETP-SMART-LEM-1, within PNCDI IV.

References

[1] H. Hassani, X. Huang and E. Silva, "Big Data and Climate Change," Big Data Cogn. Comput., 2019.

[2] S. Tatineni, "Climate Change Modeling and Analysis: Leveraging Big Data for Environmental Sustainability," *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY* , pp. 76-87, 2020.

[3] S.-S. M. Ajibade, A. Zaidi, F. V. Bekun, A. O. Adediran and M. A. Bassey, "A research landscape bibliometric analysis on climate change for last decades: Evidence from applications of machine learning," *Environment,* 2023.

[4] J. Min and H. Lee, "Utilizing Big Data Analysis in Problem Finding of Design Thinking: Deriving Meaning from Climate Change in South Korea Using LDA Topic Modeling," *Archives of Design Research,* pp. 211-221, 2023.

[5] A. D. Mauro, M. Greco and M. Grimaldi, "A formal definition of Big Data based on its essential features," *Library Review,* p. 122 – 135, 2016.

[6] M. Favaretto, E. D. Clercq, C. O. Schneble and B. S. Elger, "What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade," *PLoS One,* 2020.

[7] R. Talib, M. Kashif, S. Ayesha and F. Fatima, "Text Mining: Techniques, Applications and Issues," *(IJACSA) International Journal of Advanced Computer Science and Applications,* pp. 414-418, 2016.

[8] D. Khurana, A. Koli, K. Khatter and S. Singh, "Natural Language Processing: State of The Art, Current Trends and Challenges," *Multimedia Tools and Applications,* 2022.

[9] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications* , p. 15169–15211, 2018.

Andreea-Mihaela NICULAE – Research Assistant and PhD student at Bucharest University of Economic Studies, Romania; obtained Bachelor's Degree in Economic Cybernetics in 2020, and Master's Degree in Database Support Systems in 2022; former Erasmus+ student at Athens University of Economic and Business, attending Statistics master courses, Athens, Greece; former Erasmus+ student at Hanze University of Applied Sciences, studying IT Management, Groningen, The Netherlands.

Alin-Gabriel VĂDUVA earned his bachelor's degree in Economic Informatics in 2022 and his master's degree in Databases – Support for Business in 2024. He is currently pursuing a Ph.D., focusing on the trustworthiness of artificial intelligence algorithms in business. Professionally, he works as a Data Scientist in the banking sector, contributing to anti-money laundering projects. His research interests include mathematics, machine learning, data mining, deep learning, and generative AI.