Acta Scientific AGRICULTURE (ISSN: 2581-365X)

Volume 9 Issue 2 February 2025

Research Article

Leveraging Text Mining to Analyze Climate Change Discourse and Trends: A Computational Approach

Lakshmi Sonkusale[1]*, Vivek Kumar[2] and Himanshushekhar Chaurasia[3]

[1]*Ph.D., The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, India*

[2]*STO, ICAR-National Dairy Research Institute, Karnal, Haryana, India*

[3]*Scientist, ICAR-Central Institute for Research on Cotton Technology, Mumbai, India*

***Corresponding Author:** Lakshmi Sonkusale, ICAR-Indian Agricultural Research Institute, New Delhi, India.

DOI: 10.31080/ASAG.2025.09.1465

Received: January 16, 2025

Published: January 28, 2025

Abstract

Climate change remains one of the most urgent challenges facing the global community. As scientific research in this domain grows exponentially, analyzing large volumes of research papers, reports, and articles becomes crucial for identifying emerging trends, key discourse, and underlying patterns in the climate change field. This paper explores the application of text-mining techniques to systematically analyze and categorize the discourse surrounding climate change, leveraging computational approaches to identify trends, topics, and evolving themes from papers. Through web scraping and the extraction of 1000 paper titles from Google Scholar, spanning from 1980 to 2024, we employ a series of text mining methodologies including tokenization, lemmatization, and topic modeling to derive insights into the global research focus on climate change. The findings highlight key themes, regional disparities, and the shifts in the scientific community's priorities over time, offering valuable implications for policymakers, researchers etc. working towards climate action.

Keywords: Text Mining; Agriculture; Climate change; LDA

Introduction

Climate change has emerged as one of the most pressing global challenges, affecting ecosystems, economies, and societies worldwide. As the discourse surrounding climate change continues to grow, it generates vast amounts of textual data in the form of research articles, policy documents, news reports, and social media discussions. Analyzing this unstructured data is critical for identifying emerging trends, public perceptions, and research priorities. Text mining, a branch of natural language processing (NLP), provides powerful tools to extract meaningful insights from large textual datasets [1]. By leveraging computational approaches, text mining can uncover hidden patterns, track changes in discourse, and highlight key themes over time, enabling a deeper understanding of the evolving climate change narrative. Topic modeling categorizes articles based on content, involving the collection of documents, identification of features like words and phrases, and generation of word clusters [10]. It reveals semantic relationships among text documents and has applications in text summarization, sentiment analysis, document classification, dataset exploration, and information retrieval. [7] reviewed the effectiveness of topic modeling in discovering and documenting different categories.

Recent studies have demonstrated the utility of text mining in analyzing climate-related data. Topic modeling techniques such as Latent Dirichlet Allocation (LDA) have been used to identify dominant themes in scientific literature and policy discussions [4]. Similarly, sentiment analysis has been employed to assess public opinions and attitudes toward climate change from social media platforms [3]. This collection of studies leverages text mining and sentiment analysis to explore various aspects of climate change across multiple domains. [5], analyze over 16,000 documents from conservative think tanks, revealing a sustained increase in climate science discussions, challenging the idea that the era of science denial is over. [6], examine 35,000 climate change publications from 1990 to 2018, using Latent Dirichlet Allocation (LDA) to identify rising terms like "climate change adaptation" and declining ones like "pollution." [9] apply text mining to climate-related disclosures from Spanish financial institutions, finding a yearly increase in cli-

mate disclosures. [2] use text mining to analyze water scarcity's economic impacts and the role of institutions in addressing it. [8] conduct sentiment analysis on tweets about climate change, revealing emotional divides between supporters and opponents of scientific consensus. Topic discovery extracts semantic structures from unstructured text, enabling document classification and improved information retrieval. This study offers insights to enhance information retrieval for agricultural production, supply chains, and related fields [12,13]. Finally [11], analyzes scientific literature on precipitation patterns in India, providing insights for policymakers to address climate change's impact on agriculture. By extracting insights from published data, the research highlights gaps in current practices and offers guidance for policymakers to develop effective strategies to mitigate climate change impacts on Indian agriculture, supporting the nation's socio-economic stability. These studies collectively underscore the power of text mining in extracting valuable insights for understanding and addressing the multifaceted challenges posed by climate change.

This paper aims to explore how text mining can be leveraged to analyze climate change discourse and trends, providing a computational framework to process and interpret textual data. By integrating methods such as topic modeling and word frequency, this study contributes to understanding the key drivers, challenges, and opportunities within climate change discussions.

Methodology

Dataset description

The dataset for this study was collected through a web scraping process using Python in combination with the Beautiful Soup library. The primary objective was to curate a comprehensive collection of research paper titles related to climate change spanning the years 1980 to 2024. The scraped dataset, consisting of 1000 paper titles, forms the foundation for subsequent analysis in this study. Table 1 depicts the dataset, which comprises text data from Google Scholar, focusing on research articles related to climate change. It contains 1000 documents stored in a CSV file, with no missing values. The dataset, named "Climate Change," has three attributes (titles, year and hyperlinks), all of which are string types.

Dataset format	Text data
Source	Google Scholar
No. of dataset	1
Name of the corpus	Climate Change
No. of document/records	1000
No of attributes	3
Attributes type	String
Missing values	Nil
File type	CSV
Source link	scholar.google.com

Table 1: Dataset summary.

Data collection process

Google Scholar is a reliable repository for academic articles on climate change. Search queries employed using keywords like "climate change" to ensure comprehensive coverage of diverse topics. Filters were applied to focus on publications from 1980 to 2024, enabling both historical and contemporary perspectives. The web scraping process was executed using Beautiful Soup, a Python library for parsing HTML and XML documents. By analyzing the HTML structure of Google Scholar's search results, a custom Python script was developed to automate the extraction of paper titles, traversing multiple pages to build a robust dataset. Post-scraping, data cleaning ensured the removal of duplicate titles and exclusion of entries with incomplete information. A final manual review of a subset of titles validated their relevance to climate change. The resulting dataset comprises 1000 paper titles spanning over a century (1980-2024), reflecting interdisciplinary research across environmental science, policy, and technology. This dataset forms the foundation for text mining and analysis, facilitating the extraction of insights and trends in climate change research.

Topic generation

To generate topics from the collected dataset, we used the Latent Dirichlet Allocation (LDA) algorithm, a robust machine learning method for uncovering hidden themes in textual data. The dataset comprised 1000 research paper titles collected from Google Scholar, spanning 1980 to 2024. The process began with data pre-

processing, including cleaning (removing stop words, punctuation, and special characters), normalizing cases, tokenizing, and lemmatizing the text. A document-term matrix (DTM) was then created, where rows represented documents, columns represented unique words, and values were word frequencies or TF-IDF scores. The LDA algorithm was implemented using Python library Gensim specifying the number of topics to generate and obtain topic distributions for documents and word distributions for topics. Model optimization was performed by adjusting the number of topics, hyperparameters, and preprocessing techniques. The generated topics provided meaningful insights into the dataset, summarizing key themes, clustering related documents, and identifying research trends over time.

Results

Descriptive results

The descriptive analysis of the dataset revealed that before pre-processing, the total number of words was 8,410, with an average of 8.409 words per document, a minimum of 2 words, and a maximum of 25 words in a document. After pre-processing, the average number of words per document was reduced to 5.922, while the minimum remained at 2 words and the maximum decreased to 16 words which is shown in table 2. This reduction highlights the effectiveness of the pre-processing steps in eliminating non-informative and redundant content, streamlining the dataset for further text mining and analysis.

SN	Metric	Before Pre-Processing	After Pre-Processing
1	Total number of words	8,410	5965
2	Mean number of words per document	8.41	5.965
3	Minimum number of words per document	2	2
4	Maximum number of words per document	25	16

Table 2: Descriptive preprocessing results.

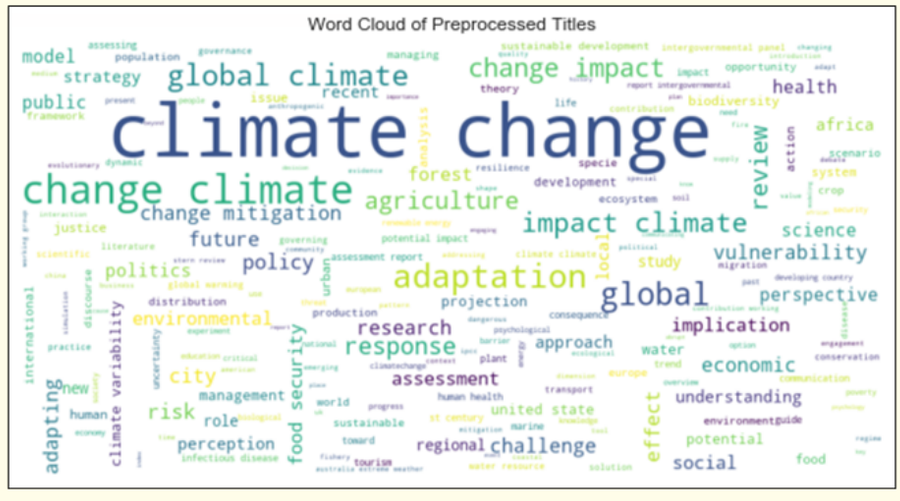

Figure 1: Word Cloud of Preprocessed Titles.

The figure 1 "Word Cloud of Preprocessed Titles," depicts the most frequently occurring words from a collection of preprocessed text data, derived from article titles. Most frequent words such as "climate", "change", "global", "adaptation", "impact", and "climate" appear in larger font sizes indicating their high frequency. Moderately frequent terms like "policy", "agriculture", "response", "vulnerability", "mitigation", "risk", and "environmental". The word cloud suggests that the dataset focuses on environmental and climate-related research, covering themes like climate adaptation, global impact, and policy considerations.

Figure 2 depicts the publication trends over time, showing the number of papers published annually. The x-axis represents the years, starting from the 1980s, while the y-axis shows the number of papers published. Initially, the number of papers was low, but from around the late 1990s, there is a sharp increase in publications. The peak is seen around 2010, with the highest number of papers published in a single year, i.e. 82. After this peak, the number of publications began to decline, indicating a decrease in the number of papers published in recent years. The blue line with markers (dots) connects the data points, visually representing this

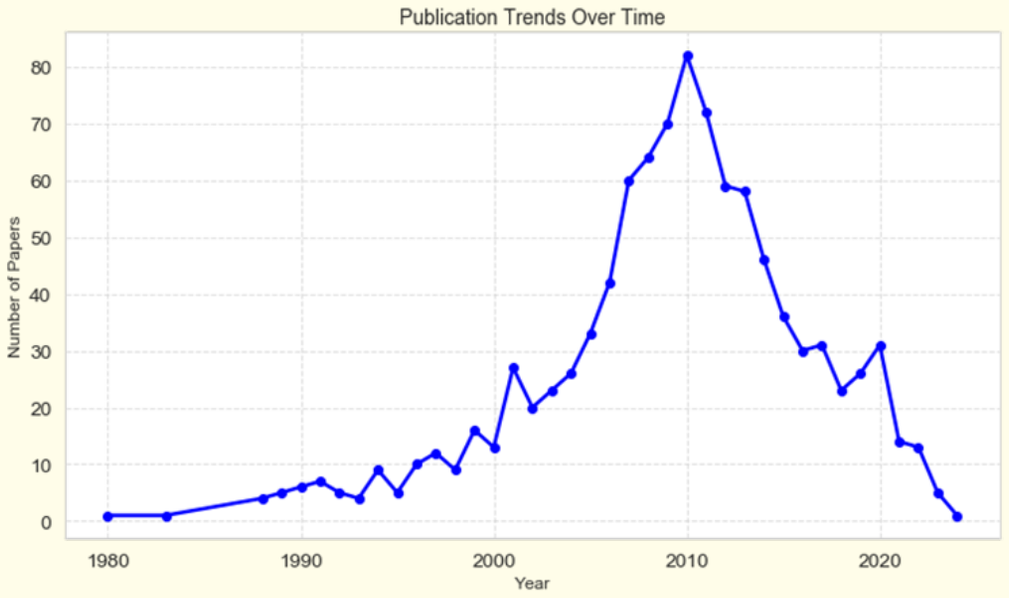

Figure 2: Publication Trend of textual data over time.

trend. This graph suggests that interest in the topic, likely related to climate change or a similar field, surged significantly around the 2000s and reached its height in the 2010.

The bar graph represents the frequency distribution of key terms extracted from a dataset likely related to climate change. The x-axis displays the most frequent words, such as "climate," "change," "global," "impacts," "adaptation," and others, while the y-axis indicates their frequencies, with values ranging from 0 to over 1000. The words "climate" and "change" are the most frequent, each appearing over 1000 times, highlighting their central importance in the dataset.

The two figures (3 and 4) illustrate the effect of text preprocessing on word frequency analysis. Figure 3 highlights the prevalence of common words such as "change", "climate", "and "of", and "the", which include stop words and case-sensitive duplicates. These frequently occurring words, however, offer little analytical value due to their grammatical nature. Furthermore, issues such as punctuation and inconsistent capitalization add to the data's noise, reducing interpretability. In contrast, figure 4, after preprocessing, presents a refined dataset where non-informative stop words have been removed, all words are converted to lowercase, and punctuation is eliminated. This preprocessing step reveals a more meaningful set of frequent terms, such as "climate", "change", "global", "impact", and "adaptation", which better reflect the dataset's themes. The comparison between these two figures clearly demonstrates the impact of preprocessing on text data. The initial dataset had noise and redundancy, while the cleaned data focused on meaningful words, enabling better insights into the text dataset, particularly for applications like topic modeling.

Figure 5 depicts the Top 10 Most Common Bigrams in the dataset in which "climate change" dominates significantly (approx. 1000 times). The remaining bigrams, such as "change climate," "global climate," "change impacts," "adaptation climate," and "impacts climate," occur much less frequently (under 200 times) and reflect key sub-themes like global impacts, adaptation, mitigation, and climate policies. The chart emphasizes that the dataset revolves around climate change and its global significance, with particular attention to impacts, mitigation efforts, and adaptation measures.

Figure 6 illustrates the Top 10 Most Common Trigrams in the dataset in which "climate change climate" and "change climate change" being the two most frequent trigrams. each exceeding 100 occurrences. Other notable trigrams include "climate change impacts," "adaptation climate change," and "impacts climate change," which focus on specific themes such as climate impacts and adaptation strategies. Additionally, terms like "impact climate change," "climate change mitigation," and "climate change adaptation" emphasize mitigation and adaptation measures as key areas of con-

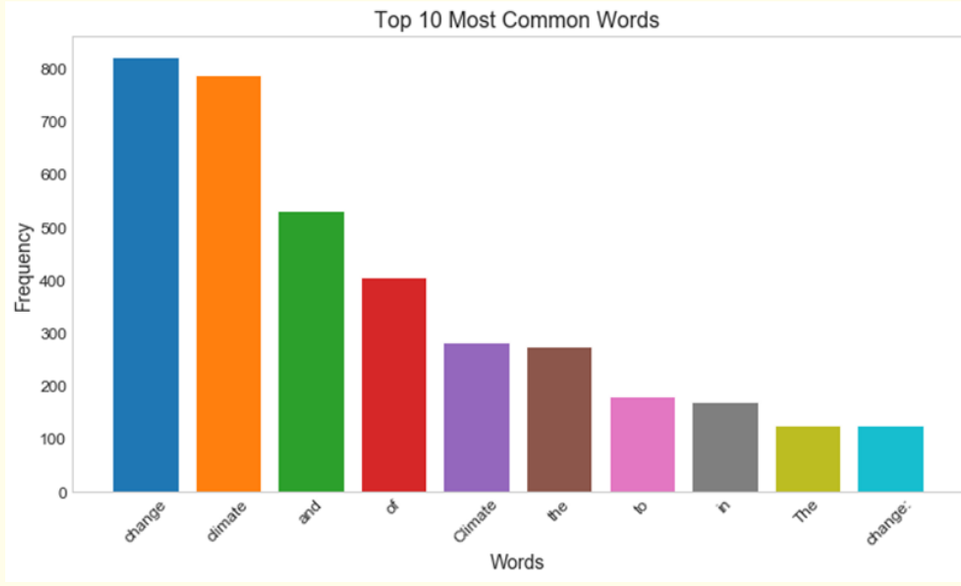

Figure 3: Most common words before preprocessing.

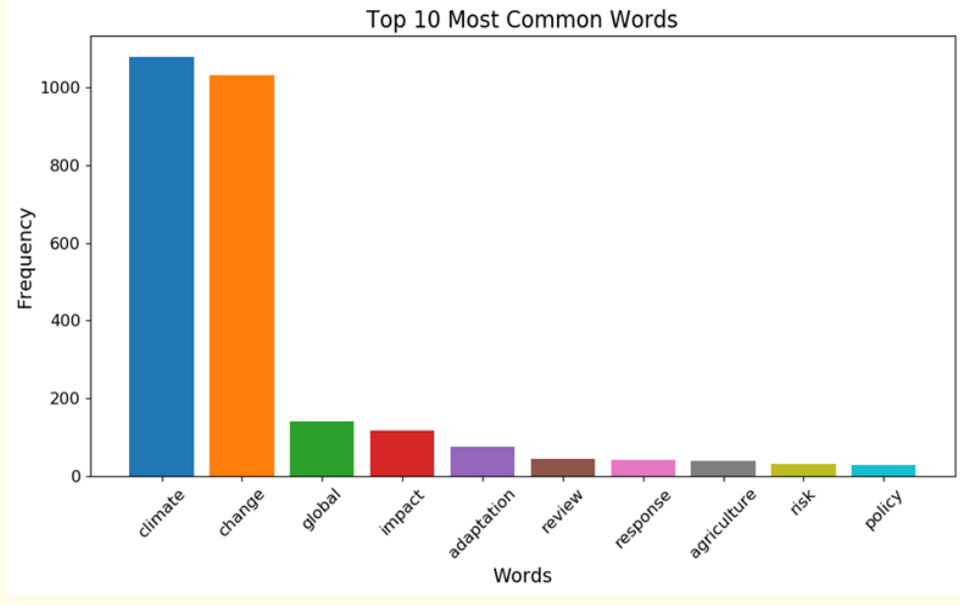

Figure 4: Most common words after preprocessing.

cern. Overall, the chart shows a clear pattern where the trigrams center on climate change as the primary topic, with strong associations to its global impacts, adaptation, and mitigation strategies, underscoring the significance of these themes in the dataset.

Topic Modeling groups documents or papers with similar themes using machine learning methods (LDA). It helps identify major research trends and focus areas in climate change literature over time. Each topic provides insights into specific subfields of climate change research, aiding scholars in understanding gaps and future directions.

The result table 3 summarizes the outcomes of the LDA topic modeling process, keywords and focus areas related to climate change research. This data represents topics and their associated keywords derived from a topic modeling analysis related to climate change research. Each topic is identified by a topic number and de-

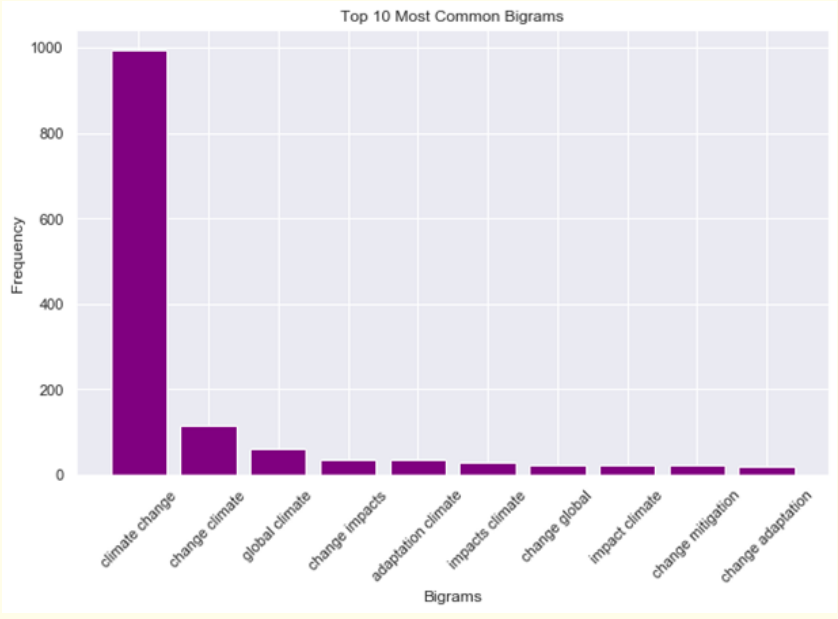

Figure 5: Bigram.

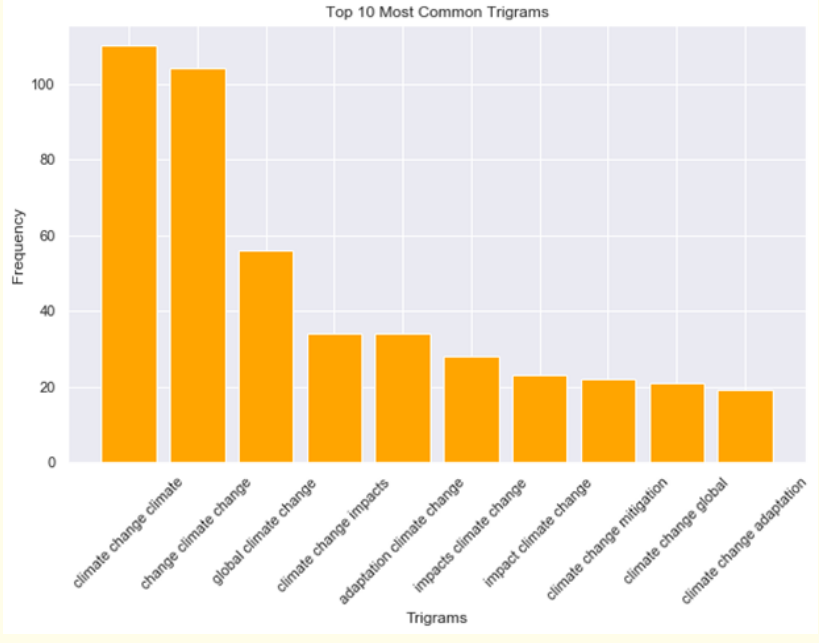

Figure 6: Trigram.

Topic No.	Top Keywords	Focus Areas
Topic 0	climate, change, adaptation, global, report, development, sustainable, assessment, developing, intergovernmental	Adaptation and development strategies
Topic 1	climate, change, global, perception, extreme, weather, policy, justice, potential, event	Global perception and policy issues
Topic 2	climate, change, response, global, impact, food, security, future, forest, disease	Responses to climate impacts
Topic 3	change, climate, review, impact, mitigation, agriculture, adaptation, global, adapting, politics	Mitigation, adaptation, and agriculture
Topic 4	climate, change, index, st, scenario, socioeconomic, reflection, local, century, research	Socioeconomic and local climate analysis
Topic 5	climate, change, global, vulnerability, discourse, review, environmental, approach, variability	Vulnerability and environmental discourse
Topic 6	change, climate, distribution, context, global, response, co, envelope, informing, specie	Global context and species distribution
Topic 7	climate, change, impact, global, science, adaptation, state, united, public, effect	Climate impacts and public science
Topic 8	change, climate, health, global, human, water, resource, opportunity, environmental, management	Health, water, and resource management
Topic 9	climate, change, impact, implication, policy, marine, adaptation, ecosystem, threat, europe	Policy implications and marine ecosystems

Table 3: Top 10 topics from "climate change" titles.

scribed by its top keywords. Each topic captures a distinct aspect of climate change. Topic 0 emphasizes adaptation and sustainable development strategies, particularly in global and developing contexts. Topic 1 focuses on global perception, extreme weather events, and policy issues, including climate justice. Topic 2 addresses responses to climate impacts, such as food security, forest health, and disease management. Topic 3 explores mitigation, adaptation, and agricultural challenges, along with political considerations. Topic 4 examines socioeconomic and local climate analyses, emphasizing regional scenarios. Topic 5 delves into vulnerability and environmental discourse, focusing on variability and review-based approaches. Topic 6 highlights the global context of species distribution and biodiversity responses. Topic 7 focuses on the impacts of climate change on public science and adaptation efforts at state levels. Topic 8 underscores the intersections of health, water, and resource management in addressing climate challenges. Finally, Topic 9 explores policy implications and threats to marine ecosystems, with a focus on Europe. Collectively, the topics provide a comprehensive overview of climate change research trends, spanning adaptation, policy, health, and ecosystem considerations.

Summary and Conclusion

The topics identified through this analysis reflect the multifaceted nature of climate change research, emphasizing adaptation, mitigation, socioeconomic impacts, biodiversity, health, and policy implications. This comprehensive categorization underscores the interconnectedness of global and local challenges in addressing climate change. The results provide a foundation for identifying research priorities and guiding policy decisions that address the diverse and critical dimensions of climate change.

Bibliography

1. Aggarwal Charu C and ChengXiang Zhai. "A survey of text classification algorithms". *Mining text Data* (2012): 163-222.

2. Aversa Dario., *et al*. "Scoping review (SR) via text data mining on water scarcity and climate change". *Sustainability* 15.1 (2022): 70.

3. Balaji K. "Text Mining in Climate Change Communication and Corporate Sustainability Reporting". Text Mining and Sentiment Analysis in Climate Change and Environmental Sustainability. IGI Global (2025): 385-418.

4. Blei David M., *et al*. "Latent dirichlet allocation". *Journal of Machine Learning Research* (2003): 993-1022.

5. Boussalis Constantine and Travis G Coan. "Text-mining the signals of climate change doubt". *Global Environmental Change* 36 (2016): 89-100.

6. Dayeen Fazle Rabbi., *et al*. "A text mining analysis of the climate change literature in industrial ecology". *Journal of Industrial Ecology* 24.2 (2020): 276-284.

7. Kherwa, Pooja and Poonam Bansal. "Topic modeling: a comprehensive review". *EAI Endorsed transactions on scalable information systems* 7.24 (2019).

8. Mi Zhewei and Hongwei Zhan. "Text Mining Attitudes toward Climate Change: Emotion and Sentiment Analysis of the Twitter Corpus". *Weather, Climate, and Society* 15.2 (2023): 277-287.

9. Moreno Angel-Ivan and Teresa Caminero. "Application of text mining to the analysis of climate-related disclosures". *International Review of Financial Analysis* 83 (2022): 102307.

10. Purver M., *et al*. "Unsupervised topic modelling for multi-party spoken discourse". In proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (2006): 17-24.

11. Ray N., *et al*. "Leveraging Text Mining for Better Policy Making to Combat Climate Change: A Bibliometric Analysis Correlating Precipitation Changes with Agriculture in India". *Annals of Library and Information Studies* 71 (2024): 435-444.

12. Sonkusale L., *et al*. "Exploring the applicability of topic modeling in SARS-CoV-2 literature and impact on agriculture". *Indian Journal of Extension Education* 22.4 (2022): 48-56.

13. Sonkusale L., *et al*. "Evaluating Text Preprocessing Methods for Discovering Quality Topics to Improve the Information Retrieval Mechanism". *Acta Scientific Computer Sciences* 5.9 (2023): 03-08.