**Extending Disaster Informatics Research: Bibliometric Dataset Expansion and Preparation for Topic Modeling (2022–2025)**

HINF 5955 Section 020 – Health Informatics Capstone Project

Department of Information Science, University of North Texas

Course Instructor: Dr. Ana Cleaveland

Submitted by Padmaja Vemula (11715789)

April 30, 2025

**Abstract**This capstone project aims to expand on previously conducted research regarding the bibliometric analysis of Disaster Informatics by collecting, cleaning, and preparing a large-scale bibliometric dataset from Web of Science, Scopus, and PubMed Central, covering the period from September 2022 to January 2025. From the three databases, 22,737 records were collected. A preprocessing workflow that comprised deduplication, language filtering, abstract validation, and publication year correction produced a collection of 19,961 articles. While topic modeling is postponed for later research, this study concentrates on dataset preparation and bibliometric analysis. The project faced several challenges, including year inconsistencies, export limitations, and collection issues, all of which were systematically resolved. This research emphasizes the necessity of selecting high-quality datasets for further advanced disaster informatics research.

**Introduction and Project Goals**

Disaster informatics is a multidisciplinary field that combines public health, data analytics, emergency management, and information science to support preparedness, response, recovery, and mitigation activities. It is more evident from the COVID-19 pandemic how important it is to have quick, accurate, and useful information during medical and humanitarian-related emergencies. As a result, studies in disaster informatics, focusing on academic discussions, technology interventions, and information-sharing during emergencies, were accelerated.

This capstone project aims to enhance the analysis by compiling and updating a dataset of publications from September 2022 through January 2025, based on the work of Tran et al. (2024), who conducted a bibliometric and topic modeling study of disaster informatics publications from 2020 to 2022. The goal was to compile an accurate and solid bibliometric dataset from PubMed Central, Web of Science, and Scopus for subsequent bibliometric analysis.

The project initially aimed to complete topic modeling and bibliometric analysis. Due to the large volume of data and cleaning challenges, the scope was narrowed. With faculty advisor approval, the focus shifted to preprocessing, bibliometric analysis, and data cleaning, preparing the dataset for future topic modeling.

Thus, the project's objectives were twofold: (1) to collect, clean, and standardize disaster informatics bibliometric data from three major databases in a methodical manner, guaranteeing high data quality and consistency; and (2) to conduct descriptive bibliometric analysis, identifying important authors, organizations, nations, and journals that contributed to disaster informatics research within the given timeframe, while establishing the framework for future related exploration through topic modeling.

**Literature Review**

Disaster informatics has evolved significantly over the past decade, gaining recognition as a multidisciplinary field central to crisis preparedness and response. Ogie and Verstaevel (2020) provided one of the earliest comprehensive overviews of the field, highlighting its foundational role in leveraging information systems to manage disaster-related data and decision-making. The COVID-19 pandemic later served as a pivotal moment in accelerating the development of data-driven emergency response technologies.

Asadzadeh et al. (2020) examined how mobile applications, big data, and GIS technologies were integrated to support emergency management during the pandemic. Complementing this, Kim et al. (2021) emphasized the value of real-time communication systems and coordinated public health data strategies for effective risk mitigation.

As disaster management strategies adapted to pandemic-specific challenges, Potutan and Arakida (2021) analyzed Japan’s evolution in emergency response practices, while Sharma et al. (2022) explored how entrepreneurial models and digital infrastructure were redefined during the crisis. On the research front, Tran et al. (2024) conducted a bibliometric and topic analysis that surfaced emerging themes within disaster informatics, reinforcing the need for clean, structured data to support advanced analyses.

Recent technological innovations have enabled deeper analysis of unstructured data sources. Grootendorst (2022) introduced BERTopic, a topic modeling framework capable of extracting coherent themes from scientific literature. Expanding on this, Mehmood et al. (2024) combined Named Entity Recognition with topic modeling to enhance disaster situational awareness from social media. Yin et al. (2024) advanced the field further by proposing CrisisSense-LLM, a large language model tailored for multi-label classification of disaster-related texts.

Collectively, these studies provide strong theoretical and methodological support for expanding disaster informatics research, validating both the bibliometric approach and the preparation for future topic modeling efforts outlined in this project.

**Methods**

This study followed a structured methodological pipeline, divided into four key stages: data collection, data preprocessing, quality assurance, and data analysis.

1. **Data Collection**  
   Articles were collected using the keywords “Disaster Informatics,” “Crisis Informatics,” and “Pandemic Crisis” across three databases: PubMed Central (PMC), Web of Science (WoS), and Scopus. From this search, a total of 22,737 articles were retrieved: 17,161 from Web of Science, 2,937 from Scopus, and 2,639 from PubMed Central. These records were merged to form the initial dataset.
2. **Data Preprocessing**  
   The merged dataset underwent multiple preprocessing steps. This included removal of non-English articles, documents without abstracts, and retracted publications. Duplicate records were eliminated using both exact and fuzzy title matching (92% similarity threshold). The dataset was also normalized (converted to lowercase) and filtered to retain only records published between September 2022 and January 2025.

A diagram of data processing

AI-generated content may be incorrect.

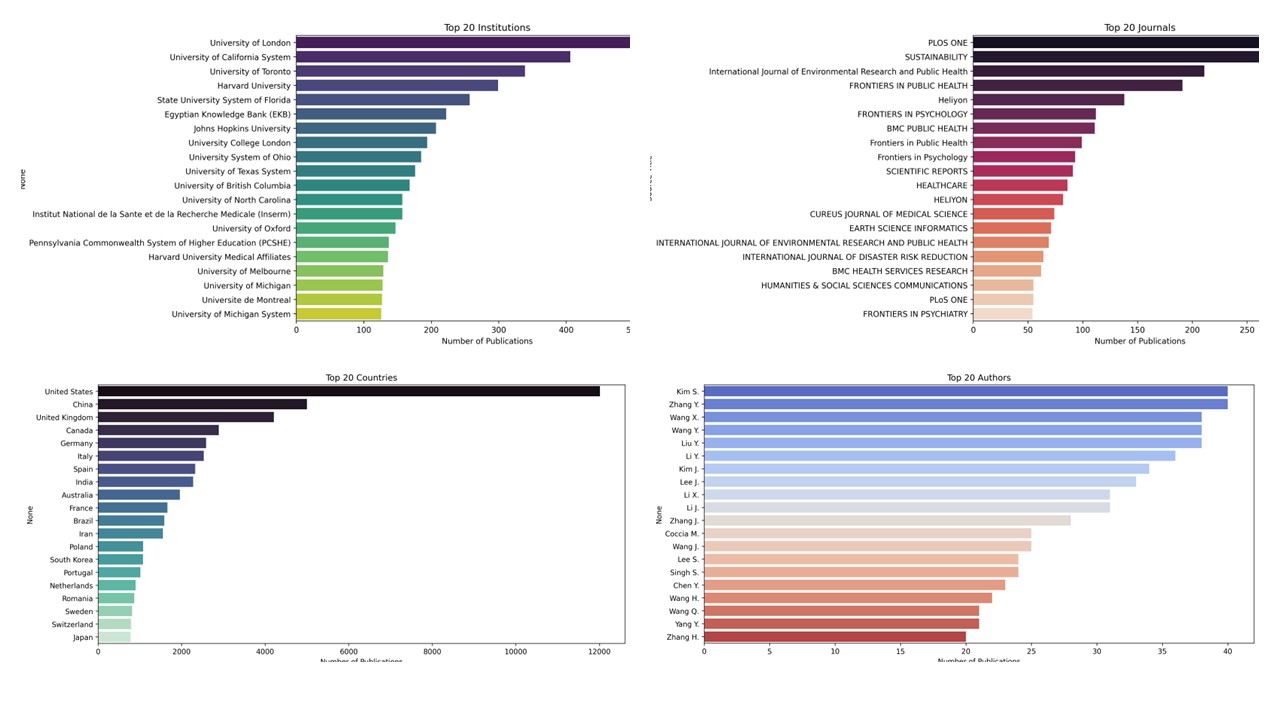
Figure 1 Research design workflow showing data collection, preprocessing, quality assurance, and analysis stages for disaster informatics publications (2022–2025)

1. **Quality Assurance**  
   After automated cleaning, a round of manual quality checks was performed. Articles missing key metadata (authors, DOIs) were reviewed. Documents deemed irrelevant to the disaster informatics domain were manually filtered out. Following these steps, the cleaned dataset was reduced from 21,744 to 19,961 high-quality entries.
2. **Data Analysis**  
   The final dataset was analyzed to identify key bibliometric trends. The top 20 countries, institutions, journals, and authors contributing to disaster informatics research during the target period were extracted. These findings formed the basis for the Results and Discussion section.

**Results and Discussion**

A detailed bibliometric analysis of disaster informatics papers from September 2022 to January 2025 was made possible by the final cleaned dataset. The United States was the country that contributed the most publications, followed by China, the United Kingdom, Canada, and Germany. These results demonstrate how North America and a few European nations continue to lead the way in disaster-related research.

The University of London was the most active provider, according to institutional data, followed by the State University System of Florida, Harvard University, the University of California System, and the University of Toronto. These universities are renowned for their multidisciplinary approach and robust infrastructure for research in informatics and public health.



**Figure 2** Top 20 countries, institutions, authors, and journals contributing to disaster informatics research

Among the researchers who published the most in this field at the author level were Kim S., Zhang Y., Wang X., Wang Y., and Liu Y. According to a review of journals, PLOS ONE was the top publication for disaster informatics, followed by Sustainability, Frontiers in Public Health, the International Journal of Environmental Research and Public Health (IJERPH), and Heliyon. The popularity of open-access, multidisciplinary publications demonstrates how widely people want to quickly make knowledge about disasters available.

These results not only confirm previous research patterns but also provide a refreshed post-pandemic view of global contributions and research dissemination. The bibliometric patterns identified in this study offer valuable insight into evolving institutional and geographic leadership, as well as preferred publication venues in the field of disaster informatics.

**Challenges**

Throughout the project, challenges such as keyword corrections, data export limitations, and publication year inconsistencies were encountered and resolved through manual validation and adaptation of processing environments.

**Future Work**

Future work will involve conducting advanced topic modeling on the prepared dataset using BERTopic and Llama 2. Comparative analysis of topic evolution between the 2020–2022 and 2022–2025 periods will also be explored.

**Summary**

The research effectively extended the bibliometric coverage of catastrophe informatics to include the timeframe of September 2022–January 2025, to sum up. Future topic modeling work was made possible by the preparation and analysis of an extensive, painstakingly cleansed dataset. This experience emphasizes how crucial it is for health informatics research to have strong data preprocessing methods.

**Capstone Project Reflection**

The knowledge and skills that I developed during the Health Informatics master's degree were effectively integrated in this capstone project. It allowed me to use my understanding of informatics tools, data preprocessing, and bibliometric analysis in a practical research environment. I was able to relate academic learning to real-world problem-solving by organizing big datasets, organizing publication records, and identifying trends in scientific collaboration.

In addition to the technical components, this project aided in my professional development. I gained knowledge about managing scope, organizing my project timeline, and avoiding last-minute pressure from this course and Dr. Cleveland's advice. These project management lessons are essential for informatics jobs in the real world. Along with focusing on the importance of career readiness, the course provided me with professional behavior guidance and interview preparation ideas, which empowered me to take on the next phase of my career.

This experience has improved my soft skills as well as my technical skills. I've developed my ability to work autonomously, troubleshoot issues, modify workflows between platforms (from Colab to VS Code), and properly record my workflow. I feel more confident going into the job market because of the capstone, which has, all things considered, served as a link between my coursework and professional health informatics practice.

**References**

Asadzadeh, A., Pakkhoo, S., Saeidabad, M. M., Khezri, H., & Ferdousi, R. (2020). Information technology in emergency management of COVID-19 outbreak. *Informatics in Medicine Unlocked, 21*, 100475. <https://doi.org/10.1016/j.imu.2020.100475>

Grootendorst, M. (2022, March 11). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv.org. <https://arxiv.org/abs/2203.05794>

Kim, D., Shinde, S. K., Lone, S., Palem, R. R., & Ghodake, G. S. (2021). COVID-19 pandemic: Public health risk assessment and risk mitigation strategies. *Journal of Personalized Medicine, 11*(12), 1243. <https://doi.org/10.3390/jpm11121243>

Mehmood, A., Zamir, M. T., Ayub, M. A., Ahmad, N., & Ahmad, K. (2024). A named entity recognition and topic modeling-based solution for locating and better assessment of natural disasters in social media. *arXiv preprint arXiv:2405.00903*. <https://arxiv.org/abs/2405.00903>

Ogie, R., & Verstaevel, N. (2020). Disaster informatics: An overview. *Progress in Disaster Science, 7*, 100111. <https://doi.org/10.1016/j.pdisas.2020.100111>

Potutan, G., & Arakida, M. (2021). Evolving disaster response practices during COVID-19 pandemic. *International Journal of Environmental Research and Public Health, 18*(6), 3137. <https://doi.org/10.3390/ijerph18063137>

Sharma, G. D., Kraus, S., Liguori, E., Bamel, U. K., & Chopra, R. (2022). Entrepreneurial challenges of COVID-19: Re-thinking entrepreneurship after the crisis. *Journal of Small Business Management*, 1–23. <https://doi.org/10.1080/00472778.2022.2089676>

Tran, N., Chen, H., Cleveland, A., & Zhou, Y. (2024). Disaster informatics after the COVID-19 pandemic: Bibliometric and topic analysis based on large-scale academic literature. *Preprint submitted to Elsevier*.

Yin, K., Liu, C., Mostafavi, A., & Hu, X. (2024). CrisisSense-LLM: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477*. <https://arxiv.org/abs/2406.15477>