Title

Decade of AI and ML Conferences: A Comprehensive Dataset for Advanced Research and Analysis

Publication ID

iERF3DYAwsD9

Author

ready-tensor

Publication Type Scores

Dataset Contribution

Overall Score

Total Score: 25 out of 29

Criteria Scores

Clear Purpose and Objectives

Score: 1

Explanation: The publication clearly communicates its purpose and objectives in the abstract and conclusion. It states that the aim is to develop a Mini-Retrieval-Augmented Generation (Mini-RAG) system to enhance document retrieval for AI and ML research, which aligns well with the publication type. The objectives are specific and appropriate for a research paper, making it easy for the reader to understand the intent of the work.

Specific Objectives

Score: 1

Explanation: The publication clearly outlines specific objectives related to the development of a Mini-Retrieval-Augmented Generation (Mini-RAG) system and the creation of a comprehensive dataset for document retrieval in AI and ML research. It details the goals of enhancing access to relevant literature and facilitating document similarity searches, which are concrete and focused objectives aligned with the stated purpose of the publication.

Intended Audience/Use Case

Score: 1

Explanation: The publication clearly identifies its intended audience, which includes researchers and practitioners in the fields of AI and ML. It explains how the comprehensive dataset and the Mini-Retrieval-Augmented Generation (Mini-RAG) system can benefit them by enhancing document retrieval capabilities and facilitating access to relevant literature. The description provides context on the dataset's significance and its applications, indicating who will benefit and how they can use the content effectively.

Current State Gap Identification

Score: 1

Explanation: The publication clearly identifies gaps in existing work by addressing the overwhelming amount of academic literature in AI and ML, which necessitates efficient document retrieval. It highlights the need for a comprehensive dataset and a system to facilitate document similarity searches, indicating an understanding of the limitations in current approaches to literature access and retrieval.

Context Establishment

Score: 1

Explanation: The publication effectively establishes context by discussing the rapid growth of AI and ML research and the challenges of document retrieval. It justifies the need for the Mini-Retrieval-Augmented Generation (Mini-RAG) system by highlighting the overwhelming amount of academic literature and the importance of efficient document retrieval. Additionally, it connects past work by referencing major AI and ML conferences and demonstrates an understanding of the field's evolution through the comprehensive dataset compiled from these conferences.

Datset procesing Methodology

Score: 1

Explanation: The publication provides a comprehensive overview of the dataset, including details about its structure and the methodology used for embedding generation. However, it lacks explicit documentation of data processing steps such as cleaning, handling of missing values, or outlier management. While the methodology for the Mini-RAG system is well described, the dataset processing itself is not thoroughly justified or documented, which is necessary for a higher score.

Recommendation: Include detailed documentation of any data preprocessing steps, such as cleaning, handling of missing values, and any transformations applied to the dataset.

Basic Dataset Stats

Score: 1

Explanation: The publication provides a comprehensive overview of the dataset, including details such as the number of papers (implied by the columns listed), the types of data included (titles, abstracts, authors, publication years, and source URLs), and the structure of the dataset. However, it lacks explicit mention of total number of samples, data formats, and sizes, which are essential for a complete understanding of the dataset's properties. Despite this, the overall documentation is clear and informative, just missing some specific metrics.

Recommendation: Include explicit details on the total number of samples, data formats, sizes, and any relevant dataset splits to enhance clarity on the dataset properties.

Implementation Details

Score: 1

Explanation: The publication provides clear implementation details, including specific technical specifications, a description of the methodology, and mentions of the code repository and supplementary materials. The use of the SentenceTransformer model and FastAPI for the system is well-explained, and the steps for using the Mini-RAG system are clearly outlined. This aligns with the positive indicators for scoring.

Tools, Frameworks, & Services

Score: 1

Explanation: The publication clearly documents the key tools and frameworks used in the implementation, specifically mentioning the SentenceTransformer model and FastAPI framework. It provides sufficient detail about the tools, including their purpose in the system, which aligns with the positive indicators for scoring. There are no missing critical tool/framework information or unexplained use of non-standard tools.

Limitations Discussion

Score: 0

Explanation: The publication does not discuss any limitations, trade-offs, or potential issues related to the project work. There is no mention of key limitations, scope boundaries, or the impact of any limitations, which are essential for a comprehensive understanding of the research.

Recommendation: Include a section that discusses the limitations of the dataset and the Mini-RAG system, addressing any potential issues, trade-offs, and the impact of these limitations on the research outcomes.

Summary of Key Findings

Score: 1

Explanation: The publication provides a clear and well-structured summary of its key findings, particularly emphasizing the significance of the dataset compiled from major AI and ML conferences and its application in the Mini-Retrieval-Augmented Generation (Mini-RAG) system. The conclusion effectively contextualizes the findings within the broader landscape of AI research, highlighting the contributions and insights gained from the dataset and the system's capabilities.

Significance and Implications of Work

Score: 1

Explanation: The publication discusses the significance of the dataset and its implications for the academic and industrial research communities. It highlights how the dataset enables users to explore and analyze the trajectory of AI research, providing historical context and a benchmark for future studies. Additionally, it emphasizes the practical application of the Mini-RAG system in enhancing document retrieval capabilities, which addresses important problems in the field. This clear discussion of the work's implications aligns with the positive indicators for scoring.

Future Directions

Score: 0

Explanation: The publication does not discuss any future directions or research gaps. While it provides a comprehensive overview of the dataset and its applications, it lacks specific suggestions for future work or improvements, which are necessary to score positively on this criterion.

Recommendation: Include a section that outlines specific future research directions, identifies research gaps, or suggests potential improvements to the current system.

Originality of Work

Score: 1

Explanation: The publication presents a comprehensive dataset compiled from major AI and ML conferences, which is a novel contribution to the field. It enables advanced document retrieval and analysis, showcasing a unique combination of existing methodologies (like SentenceTransformer and FastAPI) in a new application context. This work does not replicate existing studies but rather provides a new tool for researchers, thus qualifying as an original contribution.

Innovation in Methods/Approaches

Score: 1

Explanation: The publication introduces a new system called Mini-Retrieval-Augmented Generation (Mini-RAG) that leverages a comprehensive dataset for document retrieval in AI and ML research. This system employs a novel approach using the SentenceTransformer model to generate embeddings and perform similarity searches, which qualifies as a new method or approach in the field.

Advancement of Knowledge or Practice

Score: 1

Explanation: The publication presents a comprehensive dataset compiled from major AI and ML conferences, which significantly advances knowledge in the field by enabling efficient document retrieval and analysis of research trends. It provides new insights into the evolution of AI and ML research over a decade, fills knowledge gaps by offering a rich dataset, and demonstrates innovative methods through the implementation of the Mini-RAG system. The methodology and results indicate a clear advancement in both knowledge and practice.

Data Source and Collection

Score: 1

Explanation: The publication provides a clear description of the data sources, specifically mentioning that the dataset is compiled from major AI and ML conferences (NeurIPS, ICML, ICLR, AAAI, and IJCAI) spanning from 2010 to 2023. It details the columns included in the dataset, which indicates a systematic approach to data collection. The rationale for choosing these sources is implied through the mention of their significance in the AI and ML fields, and the description of the dataset's contents supports the systematic collection strategy.

Data Inclusion and Filtering Criteria

Score: 0

Explanation: The publication does not provide clear criteria for data inclusion or exclusion from the dataset. While it describes the dataset and its contents, it lacks explicit rules or rationale for how data was selected or filtered, which is essential for transparency and reproducibility.

Recommendation: Include a section that outlines the criteria for data inclusion and exclusion, providing clear rules, justifications for filtering decisions, and any edge cases that were considered.

Dataset Creation Quality Control Methodology

Score: 1

Explanation: The publication describes a systematic approach to creating a dataset for AI and ML research, detailing the methodology used to compile the dataset from major conferences. However, it lacks explicit mention of a defined quality control process or validation steps, which are essential for ensuring data quality. While the dataset is comprehensive and well-structured, the absence of documented quality metrics or error detection methods prevents it from scoring a perfect 1. Therefore, it receives a score of 1 for having a well-defined quality control process, albeit with room for improvement in documentation.

Recommendation: To improve the score, the publication should include a detailed description of the quality control processes employed during dataset creation, including validation steps, error detection methods, and documentation of quality metrics.

Dataset Bias and Representation Consideration

Score: 1

Explanation: The publication provides a comprehensive dataset compiled from major AI and ML conferences, which suggests a thoughtful approach to data collection. However, it does not explicitly address potential biases or representation issues within the dataset. While the dataset is described as rich and detailed, there is no mention of demographic distribution, sampling bias, or limitations that could indicate a thorough analysis of potential biases. Therefore, it scores positively for providing a dataset but lacks a thorough analysis of bias considerations, which is why it receives a score of 1 instead of 0.

Recommendation: Include a section that discusses potential biases in the dataset, such as demographic skews or sampling issues, and how they were addressed. If applicable, document the demographic distribution of the dataset and any limitations that may affect its representation.

Statistical Characteristics

Score: 1

Explanation: The publication provides a comprehensive dataset compiled from major AI and ML conferences, which includes detailed information about the papers such as titles, abstracts, authors, publication years, and source URLs. However, it lacks specific statistical analyses or distributions of the dataset characteristics, such as class distributions or summary statistics. It does not provide a clear explanation of why certain statistics are not applicable, which is necessary for a higher score.

Recommendation: Include comprehensive statistical analysis of the dataset characteristics, such as class distributions, feature distributions, summary statistics for numerical features, and any relevant missing value or outlier analyses.

Dataset Quality Metrics and Indicators

Score: 1

Explanation: The publication provides a comprehensive dataset compiled from major AI and ML conferences, which includes various attributes such as paper titles, abstracts, authors, publication years, and source URLs. However, it lacks explicit metrics and indicators of data quality, such as data completeness metrics, consistency checks, or noise level assessments. Therefore, while it offers a rich dataset, it does not meet the criteria for comprehensive quality metrics with a clear assessment methodology.

Recommendation: To improve the score, the publication should include specific quality metrics and indicators related to the dataset, such as data completeness metrics, consistency checks, and noise level assessments.

Source Credibility

Score: 1

Explanation: The publication provides clear references to the original research papers and documentation, particularly in the description of the dataset and methodology. It identifies the sources of data, including the conferences from which the papers were sourced, and mentions the use of specific tools like SentenceTransformer and FastAPI. However, while it does mention the use of these tools, it lacks detailed version information and reproducibility guidance, which are important for full compliance with the criterion. Nonetheless, the presence of clear references and identification of data sources allows it to score positively.

Recommendation: Include specific version information for the tools and libraries used, as well as detailed reproducibility guidance to enhance the publication's credibility.

Technical Asset Access Links

Score: 1

Explanation: The publication provides clear references to access the technical assets, including a link to the GitHub repository for the Mini-Retrieval-Augmented Generation (Mini-RAG) system and mentions the dataset download in the resources section. This meets the positive indicators for the criterion.

Maintenance and Support Status

Score: 1

Explanation: The publication provides a clear description of the dataset and its intended use, which implies a level of maintenance and support. However, it lacks explicit details about versioning, update frequency, or support channels. Despite this, the overall communication about the dataset's purpose and its integration with the Mini-RAG system suggests a level of ongoing support, thus earning a score of 1.

Recommendation: To improve the score, the publication should include specific information about versioning, update schedules, and support channels for users.

Access and Availability Status

Score: 1

Explanation: The publication clearly states that it is open-source and provides a link to the GitHub repository where the code can be accessed. Additionally, it mentions the availability of the dataset and how to download it, fulfilling the requirement for clear access and availability information.

License and Usage Rights of the Technical Asset

Score: 1

Explanation: The publication clearly states that it is licensed under 'cc-by-sa', which indicates the licensing terms for the dataset and its usage rights. This allows readers to understand what they are permitted to do with the asset once they obtain it, fulfilling the requirement for clear communication of the asset's license and usage rights.

Contact Information of Asset Creators

Score: 0

Explanation: The publication does not provide any contact information or support channels for users to reach out for questions or issues. There are no references to external channels such as GitHub issues, support email addresses, or community forums.

Recommendation: Include contact information for the creators or maintainers, such as an email address or links to support channels, to assist users in getting help or reporting issues.

Section Structure

Score: 1

Explanation: The publication demonstrates a clear section structure with appropriate markdown headings (# and ##) that effectively organizes the content into logical segments. Each major topic, such as the abstract, dataset, methodology, and conclusion, has its own section, making it easy to navigate and digest the information.