ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2025 September: 31(3): 273~292
http://dx.doi.org/10.13088/jiis.2025.31.3.273

Exploring Trends in Healthcare Open Source Software on GitHub: LLM-Based Topic Augmentation and Network Analysis*

Yeonjeong Kim
Department of Business Administration,
The Catholic University of Korea
(yeonjeongg12@gmail.com)

Hong Joo Lee
Department of Business Administration,
The Catholic University of Korea
(hongjoo@catholic.ac.kr)

The growing number of open-source healthcare projects on GitHub gives us useful information about new research trends and technological progress in the field. However, it is hard to analyze these repositories because the topics that users have defined are not always the same, and there are not enough of them. This study uses large language models (LLMs) to augment and standardize topics, which makes it easier to classify research topics in a more reliable way.

This study used GitHub repository metadata to do three kinds of analyses: frequency analysis to look at how topics have changed over time, semantic network analysis to look at how research areas are related to each other, and quadrant analysis of centrality to see how the roles of key topics have changed over time. The results show that there are fewer direct mentions of traditional data analysis topics, but AI-driven healthcare solutions, especially those that have to do with healthcare applications and patient management, have become more popular. The results also show that healthcare research is becoming more complicated as a result of both increasing technological specialization and interdisciplinary integration.

Keywords : Healthcare informatics, open-source software, GitHub repositories, topic augmentation, large language models

Received : September 5, 2025 Revised : September 21, 2025 Accepted : September 22, 2025
Publication Type : Fast Track Corresponding Author : Hong Joo Lee

1. Introduction

The rapid advancement of digital health technologies has resulted in a growing number of open-source healthcare-related projects. Open-source software plays a crucial role in making healthcare more accessible, equitable, and resilient (Choi et al., 2013; Paton et al., 2022) and is increasingly being developed and shared on platforms like GitHub (Shen & Spruit, 2019). GitHub, the most popular collaborative platform for software developers, relies on Git and offers extensive collaborative and social features (Varuna & Mohan, 2019).

Technological trend analyses have traditionally been conducted using academic papers and patents (Choi et al., 2011; Feng et al., 2023; Ghaffari et al.,

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A3A2A02093277)

2023; Kim & Kim, 2014; Lim et al., 2019; Park, 2018; Park et al., 2025). Additionally, GitHub repositories serve as a valuable resource for understanding current research trends, identifying emerging technologies, and assessing the evolving landscape of specific research fields (Chong et al., 2019; Cosentino et al., 2017), including healthcare informatics (Pejić et al., 2023; Shen & Spruit, 2019).

The collaborative dynamics of open-source software developments have been measured and analyzed in various ways, primarily using repository metadata (Dozmorov, 2018; Hu et al., 2016; Jarczyk et al., 2014; Pejić et al., 2023; Russell et al., 2018). Also, researchers have mainly conducted analyses by categorizing repositories based on researcher-assigned labels, user-defined topics, or repository descriptions through topic modeling (Chong et al., 2019; Shen & Spruit, 2019; Varuna & Mohan, 2019). Analyzing this large, unstructured dataset is challenging due to inconsistent user-defined metadata, varying terminology, and the absence of standardized topic classifications.

To deal with these problems, this study looked at healthcare-related repositories on GitHub in a systematic way to find patterns in research topics and technological advances. A major innovation in this study is the application of large language models (LLMs), particularly ChatGPT-4o, for topic augmentation and standardization. Traditional approaches to topic classification rely heavily on predefined repository metadata (Chong et al., 2019; Shen & Spruit, 2019; Varuna & Mohan, 2019), such as descriptions and user-assigned topics, which often lack consistency and completeness. We used LLM-based augmentation to improve topic identification and make sure that healthcare-related repositories were analyzed with more standardized topics.

The study's goal is to look at trends in open-source software projects in the healthcare field using these augmented topics. It also wants to look at how the structure of topic networks changes over time, which will help us understand how new technologies change and expand healthcare informatics. Based on these objectives, the research questions are formulated as follows.

1) How can topics of GitHub repositories be augmented with the support of LLMs?
2) What semantic structures and communities emerge from healthcare-related repositories?
3) How does the structure of topic networks change over time?

Three different analyses were used to reach these goals. We used frequency analysis to look at how repositories were distributed over time and to find changes in the importance of topics over time. We used semantic network analysis and community detection techniques to look at how research topics are related to and interact with each other. We used quadrant analysis of centrality to look at the importance and changing roles of key topics by looking at betweenness centrality, which measures how much of an influence an intermediary has, and eigenvector centrality, which measures how much of an influence an entity has.

The results showed that AI-driven healthcare solutions, especially those for medical applications and patient management, are becoming more popular. On the other hand, there have been fewer mentions of traditional data analysis topics, probably because they are being used more widely and integrated into other technologies. The semantic network analysis showed that healthcare research is moving in two directions: towards more technological specialization and towards more integration across disciplines. Also, quadrant analysis of centrality metrics showed that the roles of key technologies have changed. For example, Python, healthcare monitoring, and web applications have gone from being mostly influential entities to being critical bridging technologies facilitating interdisciplinary collaboration. The remainder of this paper details the data, methodology, results, and implications of this study.

2. Data

The metadata of GitHub repositories listed in Table 1 was collected via the GitHub API using the following healthcare-related search keywords: clinic, clinical, doctor, health, healthcare, healthtech, hospital, medical, medical automation, medicine, patient, precision medicine, and telemedicine. These keywords were identified based on prior studies (Harper, 2024; Shen & Spruit, 2019) and were expanded to include recent terms related to digital healthcare. Data collection included repositories created through 2023, resulting in a total of 28,533 repositories. The preprocessing pipeline consisted of three steps: (1) removal of duplicate repositories and those related to veterinary medicine, (2) exclusion of repositories with a size of 3 KB or less, and (3) elimination of repositories that used programming languages appearing in fewer than 10 repositories across the dataset. After preprocessing, 21,408 repositories remained for trend analysis. Figure 1 displays the distribution of repositories across each search keyword.

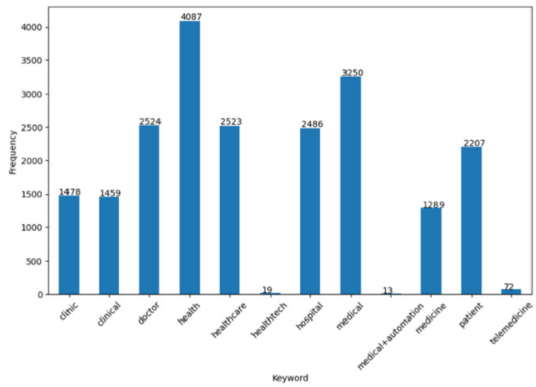

Figure 1. Distribution of Healthcare-Related GitHub Repositories by Search Keyword

3. Methods

3.1. Topic augmentation procedure

To systematically analyze topics of repositories in healthcare, we leveraged large language models (LLMs) to define topics for analysis. In this paper, we used ChatGPT-4o as our LLM for topic augmentation and identification. This approach was necessary because relying solely on repository metadata (descriptions and README files) and creator-defined topics presented several challenges. As illustrated in Figure 2, approximately 68% of the repositories either lack user-defined or share similar topics despite different research focuses.

Table 1. GitHub Repository Metadata Attributes

Attributes	Descriptions	Mean value
name	Repository name	-
url	Repository URL	-
description	Brief description of repository purpose	-
owner	Repository owner name	-
owner_type	Owner classification (“organization” or “individual”)	-
stars	Number of repository bookmarks	26.442
forks	Number of repository forks	9.577
language	Programming languages used	-
created_at	Repository creation date	-
updated_at	Latest update date	-
topics	Repository topic tags	-
open_issues	Number of unresolved issues	3.812
size	Repository size (KB)	35,168.4
readme	README file URL	-
license	Repository license type	-
contributors	Number of unique contributors	2.411
commits	Total number of commits	18.424

The average number of topics per repository is around six, yet about 85% of repositories include fewer than this average. This inconsistency showed that predefined topics could be interpreted differently across research contexts. Furthermore, repositories with similar descriptions were sometimes categorized under different topics, highlighting the lack of standardization in user-defined topics and their limited utility for systematic analysis.

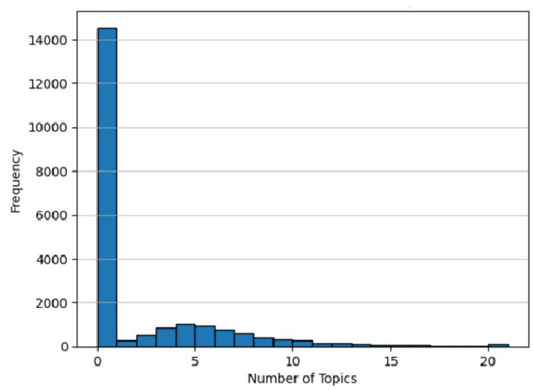

Figure 2. Distribution of User-Defined Topics per Healthcare Repository

To address these challenges, we developed a standardized topic classification system using prompt engineering with LLMs. This system synthesized multiple information sources, including repository descriptions, existing user-defined topics, and README files. We standardized the number of topics per repository to six, based on the average number of topics in repositories that had user-defined topics.

The LLM-based topic augmentation was applied using two approaches, depending on the number of existing topics:

1. For repositories with five or fewer predefined topics: The system retained relevant existing topics and generated additional topics to reach the six-topic threshold.
2. For repositories with six or more predefined topics: The system selected up to six relevant topics from the existing set. If fewer than six topics were selected, additional topics were generated to reach the threshold.
3. Topic standardization: The system changed all topics into lowercase. If both singular and plural forms existed, the system standardized them to the plural form. For terms with mixed usage such as "nlp" and "natural language processing," the system unified them using their acronyms (e.g., "nlp"). Additionally, if spaces were replaced with hyphens in topic

```python
def get_chatgpt_response_less6(name, URL, description, topics, readme):
    prompt = f"""
    I will give you GitHub repository information about healthcare.
    The first piece of information is the name of the GitHub repository, and it is as follows : {name}
    The second information is the url address of the GitHub repository, and it is as follows : {URL}
    The third piece of information is a Python list containing the topics of the GitHub repository, and it is as follows : {topics}
    The fourth piece of information is a description of the GitHub content, and it is as follows : {description}
    The last piece of information is the URL address for the detailed content on GitHub, and it is as follows : {readme}

    If you go to the URL address, you can see the detailed content about the GitHub repository.

    Your task is to define a Python list containing new topics using the given information.
    Use the five pieces of information above to define the Python list containing new topics.
    First, select the most relevant topics from the provided list, with no limit on the number of topics selected.
    Then, create enough new topics to make the total 6 by combining *selected topics* and new topics.
    When creating new topics using elements from 'topics', prohibit altering their form.
    Ensure that the new topics do not include synonyms or words with similar meanings to avoid duplication of concepts.
    new_topics must be a word.

    Example answer format : ['topic1', 'topic2', 'topic3']

    I don't need an answer other than a list. There's no need to explain.

    new_topics must consist of 6 elements. This is not optional.
    Each and every new_topic must include precisely 6 elements. No more, no less.

    """
    response = client.chat.completions.create(
    messages=[
      {"role": "system", "content": "You are a deep learning language model to help with Topic Labeling."},
      {"role": "user", "content": prompt}],
    model="gpt-4o",
    max_tokens=4096,
    n=1,
    temperature=0,
    )

    answer = response.choices[0].message.content.strip()

    return answer
```

(a) For repositories with five or fewer predefined topics

```
def get_chatgpt_response_more6(name, URL, description, topics, readme):
    prompt = f"""
    I will give you GitHub repository information about healthcare.
    The first piece of information is the name of the GitHub repository, and it is as follows : {name}
    The second information is the url address of the GitHub repository, and it is as follows : {URL}
    The third piece of information is a Python list containing the topics of the GitHub repository, and it is as follows : {topics}
    The fourth piece of information is a description of the GitHub content, and it is as follows : {description}
    The last piece of information is the URL address for the detailed content on GitHub, and it is as follows : {readme}

    If you go to the URL address, you can see the detailed content about the GitHub repository.

    Your task is to define a Python list containing new topics using the given information.
    Use the five pieces of information above to define the Python list containing new topics.

    First, select the most important topics, up to 6.
    If the selected topics are fewer than 6,
    create enough new topics to make the total 6 by combining 'selected topics' and new topics.
    When creating new topics using elements from 'topics', prohibit altering their form.
    Ensure that the new topics do not include synonyms or words with similar meanings to avoid duplication of concepts.

    Example answer format : ['topic1', 'topic2', 'topic3']

    I don't need an answer other than a list. There's no need to explain.
    """
    response = client.chat.completions.create(
    messages=[
      {"role": "system", "content": "You are a deep learning language model to help with Topic Labeling."},
      {"role": "user", "content": prompt}],
    model="gpt-4o",
    max_tokens=4096,
    n=1,
    temperature=0,
    )

    answer = response.choices[0].message.content.strip()

    return answer
```

(b) For repositories with six or more predefined topics

Figure 3. Prompts to create augmented topics via LLM

names, they were converted back to space-separated format for consistency.

Our prompt engineering methodology for ChatGPT-4o is presented in Figure 3, which illustrates the structured prompts used for topic standardization and augmentation.

3.2. Topic augmentation results

Before augmentation, the dataset contained a total of 43,225 topics, including 11,237 unique topics. After the augmentation procedure, the total number of topics increased to 128,462, with 26,287 unique topics. Although some terms were not fully processed, the number of such cases was minimal and had a negligible impact on the results. The effectiveness of this approach is demonstrated in Table 2, which presents examples of repository topics before and after the LLM-based standardization process.

As shown in Table 2, the process of topic standardization and augmentation shows two key aspects. First, when repositories had fewer topics, the LLM-based augmentation added contextually relevant topics (e.g., adding “clinical trials” and “FDA

Table 2. Examples of augmented topics

Original topics	Augmented topics
['ecg', 'healthcare', 'machine-learning']	['ecg', 'healthcare', 'machine learning', 'classification', 'data analysis', 'signal processing']
['ai', 'drugs', 'machine-lerning', 'pharmacy']	['ai', 'drugs', 'machine learning', 'pharmacy', 'clinical trials', 'fda approval']
['clinic', 'css', 'frontend', 'html', 'javascript', 'js', 'project', 'ui', 'website']	['clinic', 'frontend', 'html', 'javascript', 'ui', 'website']
['biometrics', 'blockchain-technology', 'fhir', 'healthcare-application', 'identity-management', 'master-patient-index', 'medical-informatics', 'patients']	['biometrics', 'blockchain technology', 'fhir', 'identity management', 'master patient index', 'medical informatics']

approval" for repositories related to drugs and pharmacy). Second, in cases where repositories contained more than six topics, the procedure filtered out overly general or redundant terms—such as "patient" or "healthcare application"—to improve the clarity.

After augmenting topics using ChatGPT-4o, we conducted three main analyses to understand the evolution and relationships of healthcare-related technologies in GitHub repositories: frequency analysis, semantic network analysis, and centrality-based quadrant analysis. The frequency analysis examined the temporal distribution of repositories and topic trends. The semantic network analysis, utilizing the Louvain community detection algorithm, investigated the interactions between different research fields and their development patterns over time. Finally, the quadrant analysis evaluated the roles and significance of key topics through betweenness and eigenvector centrality metrics.

4. Results

4.1. Frequency analysis

Figure 4 displays the yearly distribution of data, indicating a general rise in repository numbers over time, with a significant increase in 2020. Figure 5 illustrates the annual trends of the top 10 most frequent topics in the dataset. Although most topics exhibit a steady upward trend, the frequencies of several data analysis-related topics—such as machine learning, deep learning, medical imaging, and data analysis itself—have declined since 2020. This drop may indicate that developers are transitioning to emerging technologies or that these established technologies have reached a mature phase. Also, machine learning, deep learning, and data analysis are now widely used and often built into other research and applications, so they might not be mentioned as much anymore, which could explain why they are less common.

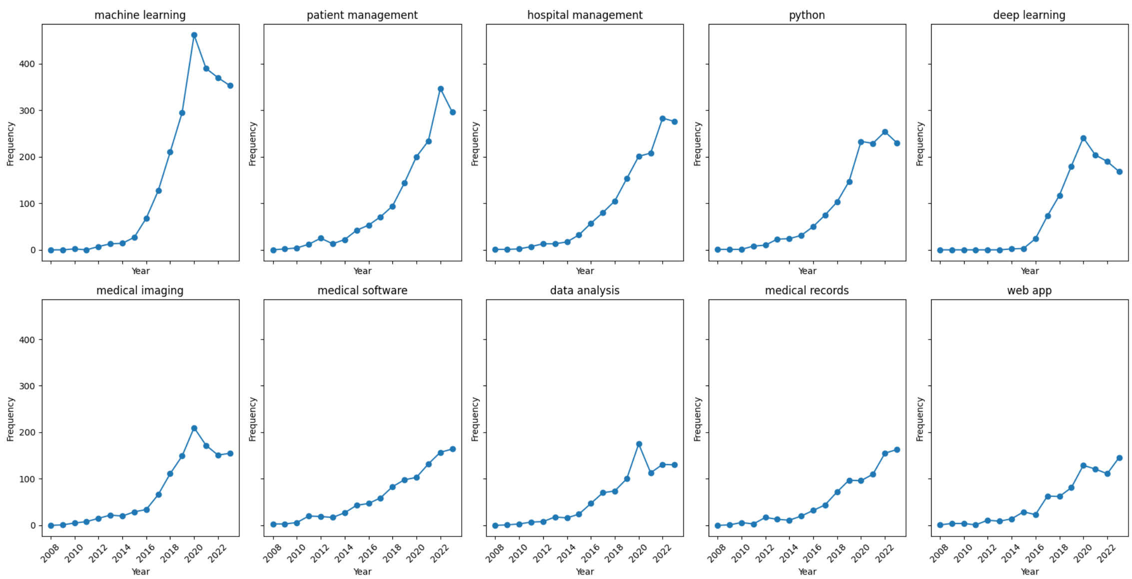

Figure 5. Annual trends of the Top 10 Most Frequent Topics

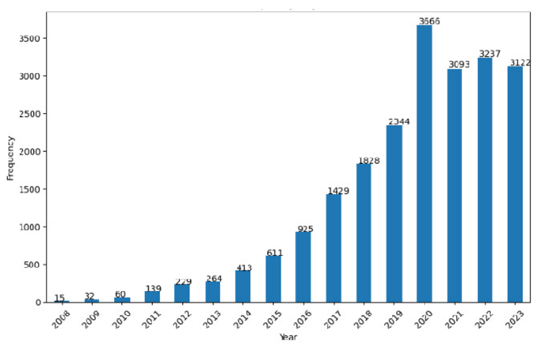

Figure 4. Yearly Distribution of Repositories

4.2. Semantic network analysis

To examine the interactions and developmental patterns across research fields, changes in the semantic network were analyzed in four-year increments. To determine the optimal network partition, we employed the Louvain Community Detection Algorithm (Blondel et al., 2008) from the NetworkX library, which uses a heuristic approach based on modularity optimization. Node size was determined based on the degree of connectivity within each community, which was identified by analyzing both the included words and the strength of their connections. The resulting semantic networks are shown in Figure 6 (a) through Figure 6 (d), and the top three central words in each community, along with the corresponding community definitions, are provided in Table 3. These community definitions were generated by ChatGPT-4o based on the topics and their centrality degrees within each community.

- 2008-2011: Early Development and Diversification

At this time, technologies related to healthcare were still in their early stages, which led to the formation of separate research communities. Medical software, health informatics, patient management, medical imaging, and automation are all fields that grew up on their own. Mobile health and telemedicine also started to become more popular, which shows how important mobile apps are becoming in healthcare. But these areas stayed mostly separate, with little integration between different fields of technology.

- 2008-2015: The Rise of Computer Vision and Systematization

As healthcare technologies advanced, research topics became more organized, and machine learning methods started to become more popular. There were big improvements in machine learning and medical

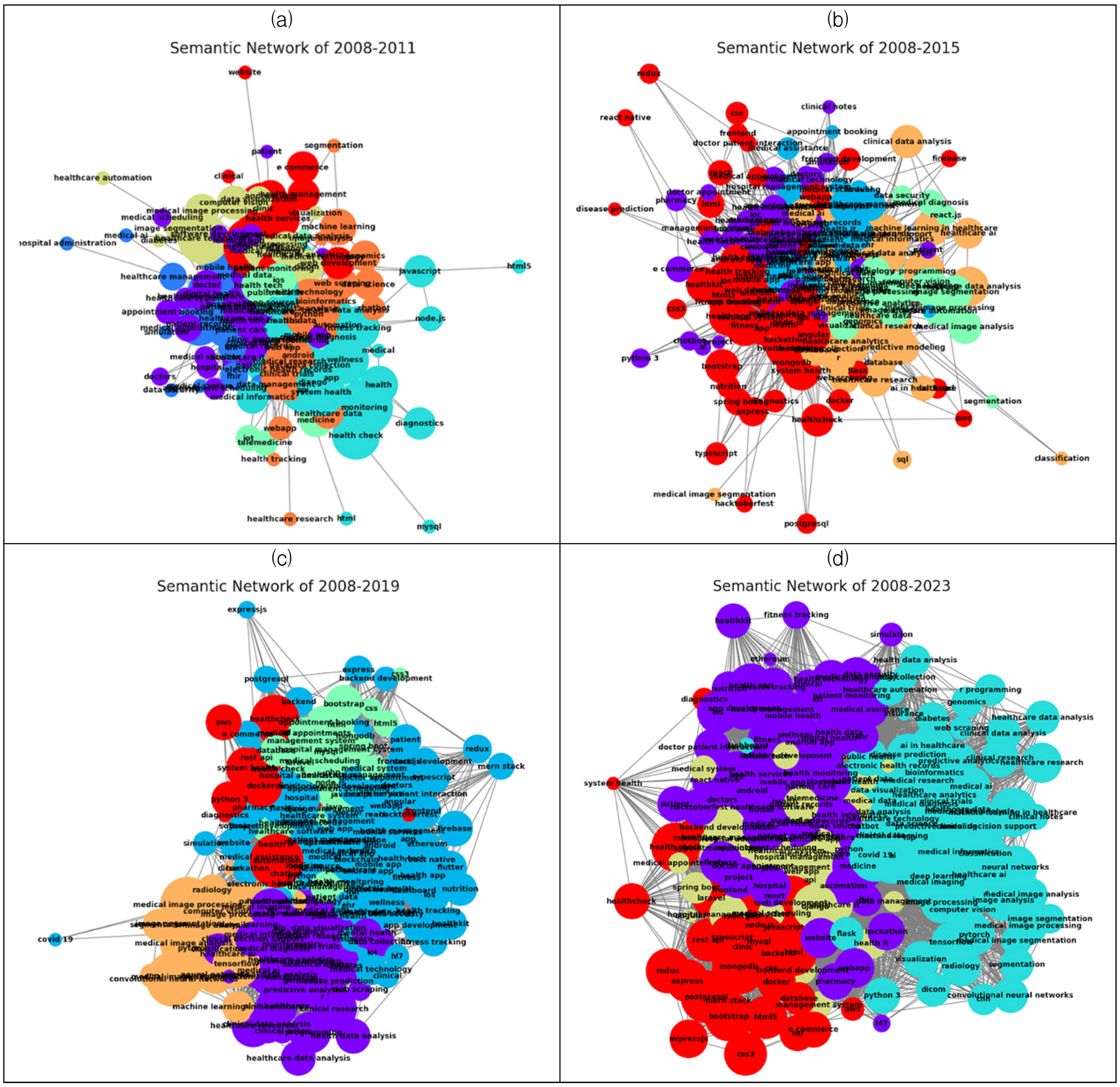

Figure 6. Evolution of Semantic Networks in Four-Year Intervals (2008-2023)

imaging. Hospital information system management and healthcare automation became important areas of interest. As time went on, data analysis techniques were applied in a wider range of areas, and there was more of a focus on visualization and machine learning. Mobile health technologies kept getting better, which made it easier for telemedicine and remote patient monitoring to become more popular.

Table 3. Community Definitions and Top Topics in the Semantic Network (2008-2023)

Yearly Grouping	Community	Top 5 Topics by Degree	Community Definition
2008-2011	Community 1	healthcare, open source, medical software, health informatics, patient management	Operations Management Systems
	Community 2	healthcare app, medical records, patient care, digital health, mobile health	Data Management and Applications
	Community 3	health monitoring, public health, patient data, web app, api	Monitoring and Data Analysis
	Community 4	android, mobile app, health technology, app, ios	Mobile Health and Telemedicine
	Community 5	medical imaging, healthcare technology, image processing, dicom	Medical Imaging and Computer Vision
	Community 6	python, data analysis, clinical trials, medical research, automation, data management	Data Analysis and Automation (Excluding Images)
	Community 7	patient monitoring, mental health, clinic, health services, web development	Patient Monitoring and Mental Health
2008-2015	Community 1	healthcare, open source, medical, java, medical app	Healthcare Applications and Automation Systems
	Community 2	patient management, web app, health informatics, medical software, hospital management	Hospital Information System Management
	Community 3	medical imaging, healthcare technology, dicom, image processing, visualization	Introduction of Computer Vision
	Community 4	python, data analysis, public health, data visualization, machine learning, medical research	Data Analysis and Visualization (Excluding Images)
	Community 5	health, health monitoring, healthcare app, mobile health, aip, javascript, telemedicine	Mobile App Development and Telemedicine
2008-2019	Community 1	python, machine learning, data analysis, medical data, data visualization, medial research	Advanced Research on Data Analysis (Excluding Images)
	Community 2	healthcare, web app, open source, medical, health	App and Web Development
	Community 3	healthcare app, hospital management, patient management, medical software, medical records	Data and System Management
	Community 4	healthcare technology, medical imaging, deep learning, dicom, healthcare ai	Advanced Computer Vision and Deep Learning
	Community 5	automation, monitoring, docker, python 3, health check, system health	Healthcare Automation and System Monitoring
2008-2023	Community 1	healthcare, open source, medical, health, medicine	Mobile App Development and Remote Management
	Community 2	python, machine learning, covid 19, healthcare technology, ai	General Data Analysis
	Community 3	healthcare app, web app, patient management, hospital management, medical software	Data and System Management
	Community 4	javascript, react, node.js, web development, react.js	Web Development and Monitoring

- 2008-2019: Expansion of Deep Learning and Advanced Automation

By this time, deep learning had become an important part of healthcare research. Machine learning and deep learning were key parts of medical imaging, patient management, and system automation. The integration of healthcare applications and hospital management systems improved data management and operational efficiency. Web and app development picked up speed as more and more people relied on digital platforms for healthcare services. Automation technologies also grew, which made it easier to keep an eye on and run healthcare systems more efficiently.

- 2008-2023: Specialization and Integration of Healthcare Technologies

Healthcare technologies have become more specialized and have also helped different fields work together better. Many people use AI-driven solutions, and mobile healthcare applications are now necessary for managing patients and providing healthcare services from a distance. The coming together of web technologies, mobile platforms, and AI has made healthcare systems that are very interconnected. Interdisciplinary collaboration has also gotten stronger, which has led to more complete and scalable healthcare solutions.

To evaluate the formation of each network's communities, partition quality was employed. This metric assesses the quality of community structures by analyzing two measures: modularity and coverage (Fortunato, 2010). Modularity indicates how strongly edges are concentrated within communities, with values closer to 1 suggesting that the community divisions accurately reflect the network structure. Coverage represents the proportion of edges contained within communities, where higher values indicate fewer connections between communities and more edges lying within them. As shown in Table 4, both modularity and coverage decrease over time, suggesting that as the network expands, inter-community connections increase, thereby blurring community boundaries.

Table 4. Partition Quality Metrics by Year Groups

Year Groups	Modularity	Coverage
2008-2011	0.5286	0.8464
2008-2015	0.4297	0.7580
2008-2019	0.4413	0.6858
2008-2023	0.4028	0.6470

The analysis revealed that while various technologies initially converged in the early stages, they gradually became more segmented and specialized over time. The distinction between analysis and development grew more evident as time progressed, yet connections among communities also strengthened, thus accelerating the construction of integrated healthcare systems.

4.3. Quadrant Analysis of Centrality

The top 20 topics were selected based on TF-IDF for all data, with search keywords removed, resulting in a final set of 18 topics. These topics are summarized in Table 5. Centrality was evaluated using two metrics: betweenness centrality and eigenvector centrality (Wasserman & Faust, 1994). Betweenness centrality quantifies how often a node appears on the shortest paths between other nodes, reflecting the degree to which a word acts as a connector. Eigenvector centrality evaluates a node's influence by accounting for both the number of its connections and the significance of the nodes it is linked to.

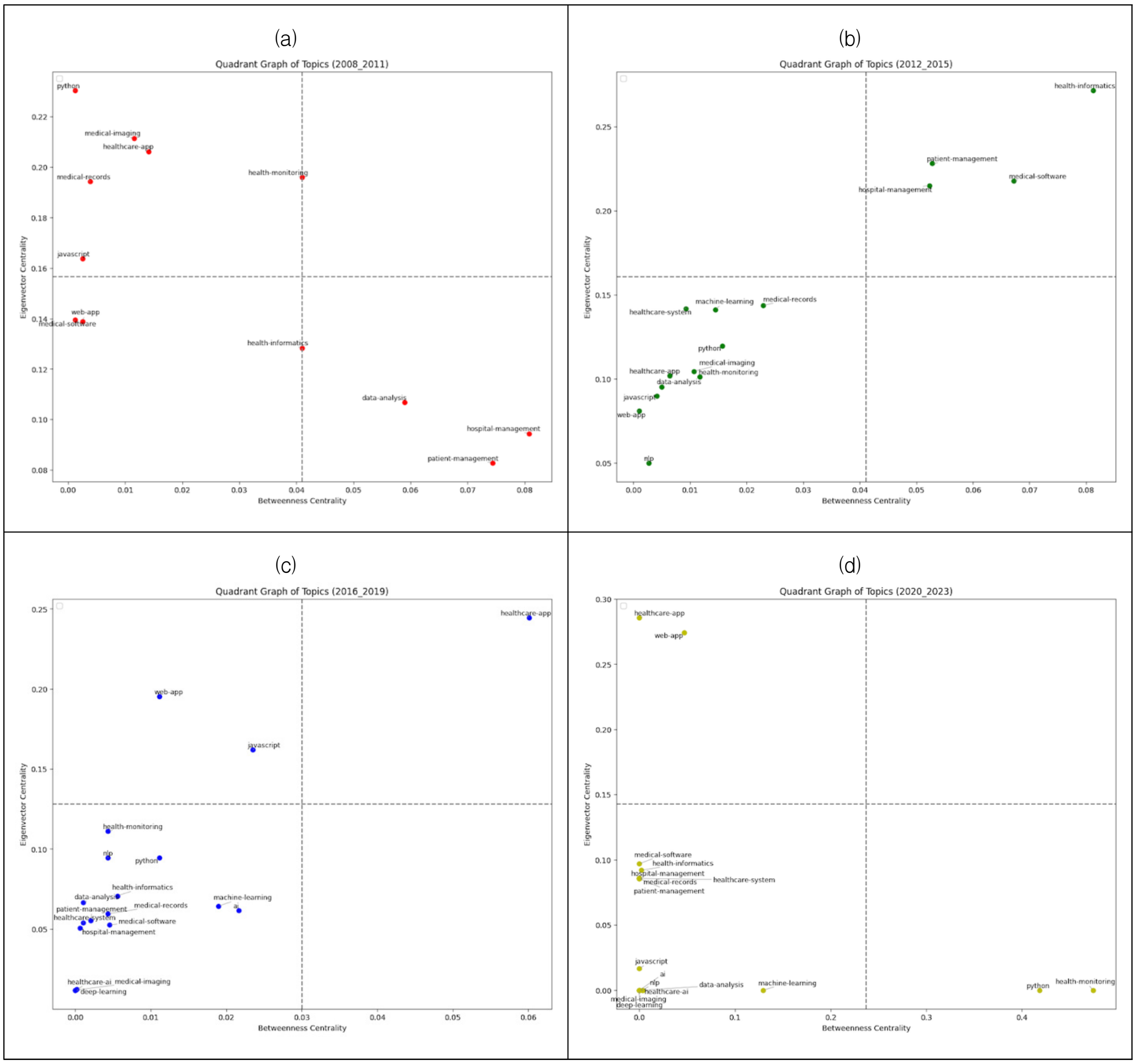

Figure 7. Centrality Changes of Topics over Time

Table 5. Topics Included in the Quadrant Analysis

ai	data analysis	deep learning	health informatics
healthcare monitoring	healthcare ai	healthcare app	healthcare system
hospital management	javascript	machine learning	medical imaging
medical records	medical software	nlp	patient management
python	web app	-	-

To clarify the role and significance of each topic, the data was divided into four-year intervals, and changes in topic centrality were examined.

The graphs in Figure 7(a) through 7(d) are divided into quadrants based on the median values of betweenness centrality (intermediary roles) and eigenvector centrality (influence), enabling a comparison of each topic's role in the network. Topics in the second quadrant (low intermediary roles, high influence) include, for the 2008–2011 year group, Python, medical imaging, healthcare app, health monitoring, web app, and JavaScript; for 2012–2015, health informatics, patient management, medical software, and hospital management; and for 2016–2019, healthcare app and medical records. In contrast, the fourth quadrant (high intermediary roles, low influence) contains Python and health monitoring for the 2020–2023 year group.

Next, an integrated analysis of community definitions and centrality changes was conducted. Python has consistently been part of data-analysis-focused communities. During 2008–2011, it showed low intermediary roles but high influence in the "Data Analysis and Automation (Excluding Images) Community." However, in 2020–2023, within the "General Data Analysis Community," it exhibited the opposite trend.

Medical imaging was part of the "Medical Imaging and Computer Vision Community" in 2008–2011, and healthcare app showed low intermediary roles but high influence in both the "Data Management and Applications Community" of 2008–2011 and the "Data and System Management Community" of 2016–2019. Health monitoring, web app, and JavaScript also demonstrated low intermediary roles but high influence in the "Monitoring and Data Analysis Community" of 2008–2011, while healthcare monitoring shifted to a high intermediary role in the "Mobile App Development and Remote Management Community" of 2020–2023.

Also, health informatics, patient management, medical software, hospital management, and medical records had low intermediary roles but high influence in the "Hospital Information System Management Community" from 2012 to 2015.

This analysis of centrality changes over time using a quadrant graph showed that as the network got bigger and more complicated, some topics' centralities went down. However, recent trends show that eigenvector centrality is rising for health informatics, healthcare system, medical software, medical records, hospital management, patient management, healthcare apps, and web apps. This means that they have more overall influence. It is important to note that web app's betweenness and eigenvector centrality both went up, which shows

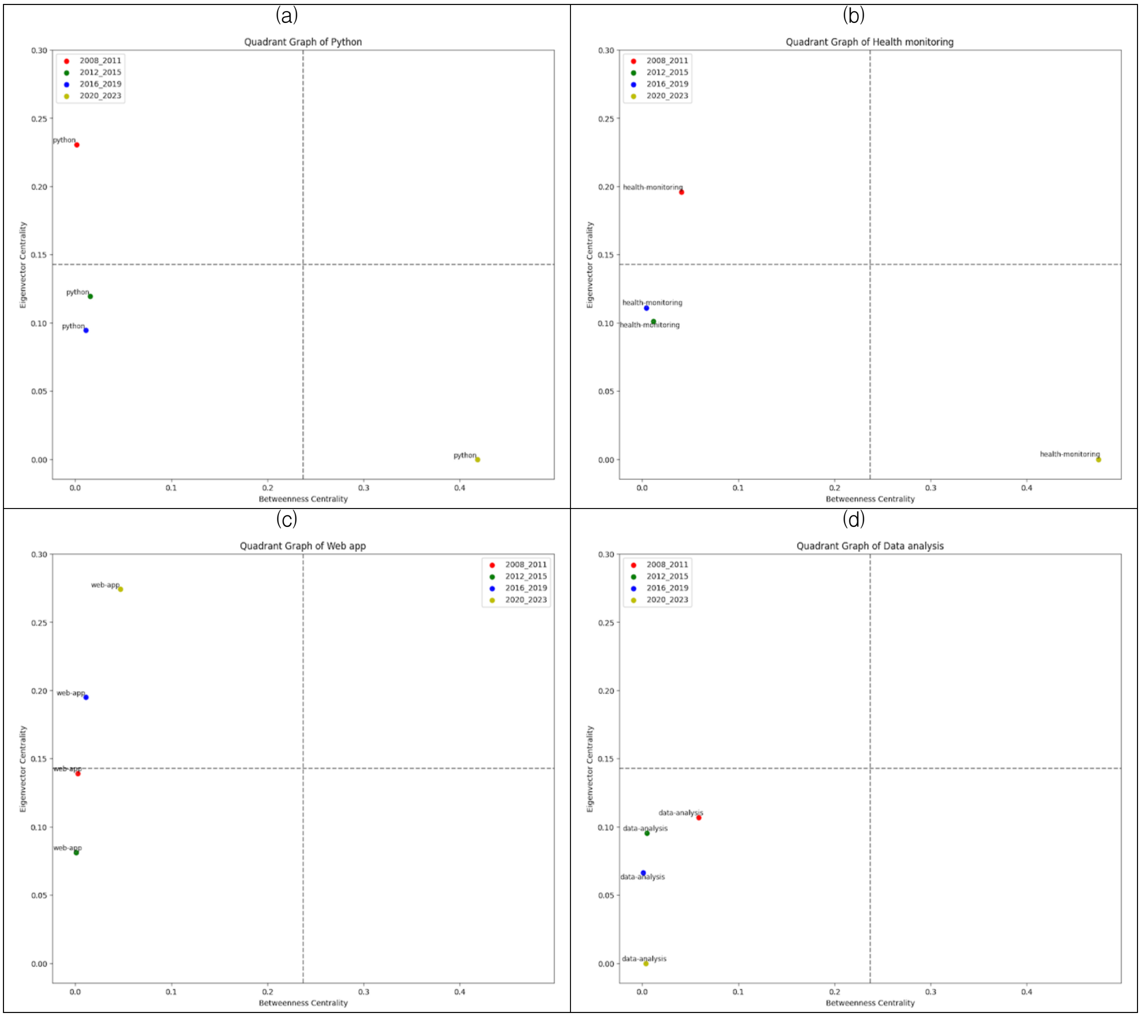

Figure 8. Centrality Changes over Time for Python, Health monitoring, Web app and Data analysis

how important it is for bringing together and managing medical data. The betweenness centrality of health monitoring, data analysis, Python, and machine learning has also gone up recently, showing that they are now better at connecting different areas of healthcare.

Overall, Python's evolution—from being more important in early data-analysis communities to being a bridge in recent years—reflects how widely it is now used in many areas of healthcare. In the same way, healthcare monitoring used to be a major factor, but now it serves more as a link between telemedicine and mobile healthcare systems. These patterns show how healthcare technologies are always changing and how information flow and connectivity are changing within and between research fields.

Web apps, on the other hand, have become very important and powerful tools for coordinating medical data and interactions.

5. Discussions

This study looked at healthcare-related GitHub repositories in a systematic way to find research trends and see how the role of important technologies has changed over time. We used LLMs to augment and standardize topics, which helped us fix the problems with user-defined topics. This made trend analysis more reliable and organized. The combination of frequency analysis, semantic network analysis, and centrality-based quadrant analysis gave us useful information about how healthcare research and technology have changed over time.

We did centrality analysis by breaking the dataset into four-year chunks, and we did semantic network analysis by adding four-year chunks of data one at a time. This method was chosen to deal with the fact that older data is becoming less important compared to newer data as the amount of recent data grows quickly. This makes the impact of newer data stand out more. Even though the methods used in the two analyses were slightly different, they both produced similar results that accurately showed changes over time.

The frequency analysis showed a substantial rise in healthcare-related repositories over time, with a marked increase in 2020, probably due to the COVID-19 pandemic. Most topics grew steadily, but traditional data analysis-related topics like machine learning, deep learning, and medical imaging saw a drop in explicit mentions after 2020. This could mean that they are moving from new research areas to established methods that are now used in a wider range of applications. This change shows how important AI-driven healthcare solutions are becoming. They now go beyond research and into real-world use in medical software, healthcare systems, and patient management.

Semantic network analysis shows that healthcare research is moving in two main directions: more technological specialization and better integration between fields. At first, healthcare technologies were split up into separate research groups, like medical imaging, patient management, and data analysis. As time has gone on, these areas have come together, especially with the rise of AI and automation. The Louvain community detection method shows that the community structure is changing from separate domains to research areas that are connected. This is especially true for AI-driven applications in healthcare informatics, mobile health, and automation. Mobile and web-based healthcare applications have become more popular, which shows how healthcare services are changing in the digital age.

Quadrant analysis of centrality metrics gave us more information about how the roles of important technologies are changing. In the past, Python, medical imaging, and web applications were important topics that were at the heart of healthcare research. Over time, Python, healthcare monitoring, and web apps have gone from being very important topics to being key bridging technologies with higher betweenness centrality. This indicates that they are now playing a more critical role in connecting different technological areas and supporting interdisciplinary collaboration in digital healthcare. At the same time, AI-related topics have become more important across the network, which suggests that they are more widely used and are the basis for new healthcare technologies.

Also, the LLM-based topic augmentation method used in this study worked well to get around the problems caused by inconsistent user-defined metadata. We were able to create a more complete and organized classification of healthcare-related repositories by standardizing and augmenting the topics in the repositories with an LLM. As discussed in Methods section and shown in Table 2, this process generated around twice as many topics as the user-defined ones and produced more relevant topics. The improved consistency of topics enhanced the accuracy of frequency trends, semantic relationships, and centrality measures.

6. Conclusions

This study shows the ongoing evolution of healthcare informatics, with increasing adoption of AI and data-driven approaches in real-world healthcare applications. The declining popularity of traditional data analysis topics does not diminish their importance; rather, it indicates their broader, more generalized application. On the other hand, the growth of healthcare AI, patient management tools, and web-based platforms shows that these fields will keep having an impact on the future of healthcare technologies. The combination of technological specialization and interdisciplinary integration suggests that future research will focus on refining AI applications, improving system interoperability, and enhancing digital healthcare infrastructure.

This study gives us some useful information, but there are certain limitations to consider. First, using GitHub repository metadata may lead to biases because the documentation may not be complete or consistent. LLM-based topic augmentation made it easier to find and sort topics, but some repositories may still not have enough descriptive information, which could make topic assignments and trend analysis less reliable. Second, the validation of the topic augmentation results is explanatory rather than numerical, as there is no definitive set of correct topics for each repository. Instead, the evaluation relies on the qualitative discussion provided after Table 2. As shown in Table 2, the outcomes of topic augmentation are promising, since the added topics are relevant and the excluded topics are redundant. Third, this study is mostly about repository metadata and topic trends, not about how these technologies are used or what effect they have in the real world. Centrality and network analyses can help us understand how the research landscape is changing, but they do not directly measure how well these technologies work or how useful they are in real healthcare settings. To get a better picture of new technologies, future research should use more data sources, like academic journals, clinical trial databases, or healthcare systems. By linking these data sources with open-source development trends, researchers can assess whether emerging technologies are being tested and applied in clinical practice and academic research.

This study lays the groundwork for keeping track of how healthcare technologies are changing over time by using advanced computational methods and large-scale open-source data.

References

[국내 문헌]

Choi, J. H., Kim, H. S., & Im, N. G. (2011). Keyword Network Analysis for Technology Forecasting. *Journal of Intelligence and Information Systems*, *17*(4), 227-240. https://doi.org/10.13088/jiis.2011.17.4.227

Chong, J., Kim, D., Lee, H. J., & Kim, J. W. (2019). A Study on the Development Trend of Artificial Intelligence Using Text Mining Technique: Focused on Open Source Software Projects on Github. *Journal of Intelligence and Information Systems*, *25*(1), 1-19. https://doi.org/10.13088/JIIS.2019.25.1.001

Kim, D. S. & Kim, J. W. (2014). Research Trend Analysis Using Bibliographic Information and Citations of Cloud Computing Articles: Application of Social Network Analysis. *Journal of Intelligence and Information Systems*, 20(1), 195-211. https://doi.org/10.13088/jiis.2014.20.1.195

Park, J. Y. (2018). Trend Analysis of Artificial Intelligence Technology Using Patent Information. *Journal of the Korea Society of Computer and Information, 23*(4), 9-16.

[국외 문헌]

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Choi, J., Choi, J., Lee, S. H., Hwangbo, H., Lee, I., & Kim, J. (2013). The Reinforcing Mechanism of Sustaining Participations in Open Source Software Developers: Based on Social Identity Theory and Organizational Citizenship Behavior Theory. *Asia Pacific Journal of Information Systems*, *23*(3), 1-23. http://doi.org/10.14329/apjis.2013.23.3.001

Cosentino, V., Izquierdo, J. L. C., & Cabot, J. (2017). A Systematic Mapping Study of Software Development With GitHub. *IEEE Access*, *5*, 7173-7192. https://doi.org/10.1109/ACCESS.2017.2682323

Dozmorov, M. G. (2018). GitHub Statistics as a Measure of the Impact of Open-Source Bioinformatics Software [Perspective]. *Frontiers in bioengineering and biotechnology*, *Volume 6 - 2018*.

Feng, L., Zhang, H., Wang, J., Lin, K.-Y., & Li, J. (2023). Innovation technology opportunity identification of civil aircraft mechanical connections based on generative topographic mapping. *PLoS One*, *18*(10), e0293309. https://doi.org/10.1371/journal.pone.0293309

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75-174. https://doi.org/10.1016/j.physrep.2009.11.002

Ghaffari, M., Aliahmadi, A., Khalkhali, A., Zakery, A., Daim, T. U., & Yalcin, H. (2023). Topic-based technology mapping using patent data analysis: A case study of vehicle tires. *Technological Forecasting and Social Change*, *193*, 122576. https://doi.org/10.1016/j.techfore.2023.122576

Harper, J. R. (2024). Automated Extraction and Maturity Analysis of Open Source Clinical Informatics Repositories from Scientific Literature. arXiv, https://doi.org/10.48550/arxiv.2403.14721.

Hu, Y., Zhang, J., Bai, X., Yu, S., & Yang, Z. (2016). Influence analysis of Github repositories. *SpringerPlus*, *5*(1), 1268. https://doi.org/10.1186/s40064-016-2897-7

Jarczyk, O., Gruszka, B., Jaroszewicz, S., Bukowski, L., & Wierzbicki, A. (2014). GitHub Projects. Quality Analysis of Open-Source Software. In L. M. Aiello & D. McFarland (Eds.), *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings* (pp. 80-94). Springer International Publishing. https://doi.org/10.1007/978-3-319-13734-6_6

Lim, H.-A., Vy, T. D. P., & Choi, J. (2019). Detecting Knowledge structures in Artificial Intelligence and Medical Healthcare with text mining. *Asia Pacific Journal of Information Systems*, *29*(4), 817-837. https://doi.org/10.14329/apjis.2019.29.4.817

Park, Y., Lim, S., Gu, C., Syafiandini, A. F., & Song, M. (2025). Forecasting topic trends of blockchain utilizing topic modeling and deep learning-based time-series prediction on different document types. *Journal of Informetrics*, *19*(2), 101639. https://doi.org/10.1016/j.joi.2025.101639

Paton, C., Braa, J., Muhire, A., Marco-Ruiz, L., Kobayashi, S., Fraser, H., Falcón, L., & Marcelo, A. (2022). Open Source Digital Health Software for Resilient, Accessible and Equitable Healthcare Systems. *Yearbook of Medical Informatics*, *31*(1), 67-73. https://doi.org/10.1055/s-0042-1742508

Pejić, N., Radivojević, Z., & Cvetavnović, M. (2023). Analyzing the Impact of COVID-19 on GitHub Event Trends. *Sustainability*, *15*(19), 14622. https://doi.org/10.3390/su151914622

Russell, P. H., Johnson, R. L., Ananthan, S., Harnke, B., & Carlson, N. E. (2018). A large-scale analysis of bioinformatics code on GitHub. *PLoS One*, *13*(10), e0205898. https://doi.org/10.1371/journal.pone.0205898

Shen, Z., & Spruit, M. (2019). A Systematic Review of Open Source Clinical Software on GitHub for Improving Software Reuse in Smart Healthcare. *Applied Sciences*, *9*(1), 150. https://doi.org/10.3390/app9010150

Varuna, V. T., & Mohan, A. (2019). Trend Prediction of GitHub using Time Series Analysis. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815478

국문요약

GitHub 기반 헬스케어 오픈소스 소프트웨어 동향 탐구: LLM 기반 주제 증강과 네트워크 분석

김연정* · 이홍주**

GitHub에서 오픈소스 헬스케어 프로젝트가 증가함에 따라 이 분야의 새로운 연구 동향과 기술적 발전에 대한 유용한 정보를 얻을 수 있다. 그러나 사용자들이 정의한 주제가 항상 동일하지 않고 충분하지 않기 때문에 이러한 저장소를 분석하는 것은 쉽지 않다. 본 연구는 대규모 언어 모델(LLM)을 활용하여 주제를 보완하고 표준화함으로써 연구 주제를 보다 신뢰성 있게 분류할 수 있도록 하였다. 본 연구는 GitHub 저장소 메타데이터를 활용하여 세 가지 분석을 수행하였다. 첫째, 주제가 시간에 따라 어떻게 변화했는지를 살펴보기 위한 빈도 분석, 둘째, 연구 영역 간의 관계를 탐구하기 위한 의미 네트워크 분석, 셋째, 주요 주제의 역할이 시간에 따라 어떻게 변화했는지를 살펴보기 위한 중심성 사분면 분석이다. 분석 결과, 전통적인 데이터 분석 주제에 대한 직접적인 언급은 줄어든 반면, 특히 헬스케어 애플리케이션 및 환자 관리와 관련된 AI 기반 헬스케어 솔루션이 더욱 주목받고 있는 것으로 나타났다. 또한 헬스케어 연구는 기술적 전문화 심화와 학제 간 통합의 결과로 점점 더 복잡해지고 있음을 보여준다.

주제어 : 헬스케어 인포매틱스, 오픈소스 소프트웨어, GitHub 저장소, 주제 보완, 대규모 언어 모델

논문접수일 : 2025년 9월 5일 논문수정일 : 2025년 9월 21일 게재확정일 : 2025년 9월 22일
원고유형 : Fast Track 교신저자 : 이홍주

* 가톨릭대학교 경영학과
** 교신저자: 이홍주
가톨릭대학교 경영학과
(14662) 경기도 부천시 원미구 지봉로 43
Tel: +82-2-2164-4009, Fax: -82-2-2164-4280, E-mail: hongjoo@catholic.ac.kr

저자 소개

김연정
현재 가톨릭대학교 경영학과 학사과정 재학 중이다. 주요 관심분야는 데이터 분석, 텍스트 마이닝, 마케팅 등이다.

이홍주
현재 가톨릭대학교 경영학과 교수로 재직 중이다. KAIST 산업경영학과를 졸업하고, KAIST 테크노경영대학원에서 석사 및 박사학위를 취득하였다. 주요 관심분야는 데이터 분석, 지능형 정보시스템, 온라인 사용자들의 상호작용, 헬스케어, 소셜 미디어 등이다.