AI-Based Healthcare Research Insights System

Student Names: Gulab Shanaaz Shaik Mohammad, Jai Raj Yadav

Course Title: MIS 6349 – Digital Consulting Project

Instructor: Ramesh Satya Sree Venkata Garapaty

University: University of Texas at Dallas

Date: May 11, 2025

EXECUTIVE SUMMARY

Our project develops artificial intelligence based system that will automate the literature analysis, hence addressing the increasing difficulty of manually examining huge volumes of healthcare research. Our technology uses artificial intelligence and natural language processing (NLP) that extracts text from research papers, summarizes the important points, identify the main themes, and show the findings through an interactive dashboard. Our aim was to lower the manual labor and increase the availability of study results for decision-makers and medical practitioners.

Among several techniques studied were BERTopics, LDA, PyPDF2, spaCy, BERT-based summarizers, and Metabase. LDA was chosen for topic modeling due to dataset constraints; Metabase was visualized via Docker integration. Adopting open-source, AI-driven technologies in research environments is the last advice meant to enable scalable, effective, and economical analysis of unstructured data.

INTRODUCTION

In a time when academic research is growing quickly, sifting through many articles to find important insights has become rather difficult. Manual reviewing large PDFs is time-consuming and ineffective for clinicians, researchers, and legislators as well. Designed to solve this difficulty by automating the study of medical literature using artificial intelligence and natural language processing (NLP), the AI-Based Healthcare Research Insights System Developing a complete end-to--end pipeline that could extract text from research papers, summarize major findings, uncover thematic trends, and show the results via an interactive dashboard was the main objective.

Overview of the Technology Domain:

This system converts unorganized academic data into organized, easily available insights, therefore enabling professionals to concentrate more on decision-making than on data analysis. The project shows how fast modern consulting solutions can be prototyped using freely available technologies by combining open-source, off-the-shelf tools including PyPDF2 for text extraction, spaCy for preprocessing, BERT for summarizing, LDA for topic modeling, and Metabase for visualization.

Structure of the Paper:

This work is organized as follows: it starts with a review of the tools at hand for automated research analysis then delves deeply into a few chosen technologies applied in the project. Along with important conclusions and last suggestions, a reasonable use case is given to show useful applicability. The study ends with considering technical difficulties encountered during implementation and the fixes used to get past them

TOOL LANDSCAPE & SHORTLIST

Many tools were assessed depending on their capacity in text extraction, natural language processing (NLP), summarizing, topic modeling, and visualization in order to automate the examination of healthcare research publications. The choices concentrated on scalable, open-source, integrable tools fit for usage in a local dashboarding system such as Metabase and a cloud environment such as Google Colab. The five main instruments regarded below are:

1. PyPDF2:

A Python open-source tool for text reading and extraction from PDF files. Its simplicity, fit with Colab, and efficiency in managing unstructured text input forms helped to justify it.

2. spaCy:

Preprocessing chores including tokenizing, stopword removal, and normalizing a quick and strong NLP library. The lightweight en_core_web_sm model of spaCy let effective processing in settings with limited resources.

3. BERT (from bert-extractive-summarizer):

This transformer-based summarizing tool finds the most pertinent lines from a given document by means of pre-trained BERT models. For succinct, accurate summaries of voluminous medical material, it was perfect.

4. BERT Topic:

A sophisticated topic modeling tool comprising HDBSCAN for clustering, UMAP for dimensionality reduction, and transformer embeddings. Though strong, it was not appropriate for tiny datasets and finally LDA superseded it.

5. Metabase:

A freely available business intelligence application designed for interactive dashboard creation. To see summaries, subject distributions, and keyword patterns, it was hosted locally using Docker and linked to the SQLite database.

Tool Comparison Table:

Tool	**Key**	**Pricing**	**Complexity**	**Applied in Use Cases**
PyPDF2	PDF text extracting with basic API	Free (MIT)	Low	Document parsing
spaCy	NLP preparation, quick and effective	Free (MIT)	Average	Text cleaning and tokenizing
BERT Summarizer	preparation, quick and effective	Free (MIT)	High	Text cleaning and tokenizing
BERTopic	Topic modeling with embeddings and clustering	Free (MIT)	High	Advanced topic discovery (more extensive datasets)
Metabase	Creation of dashboard, SQL support, interactive filters	Free (AGPL)	Medium	Reporting and data visualization

Standards of Evaluation Applied:

The following standards helped to rank the tools:

1. Capacity to manage fundamental chores (extraction, preprocessing, summarizing, topic modeling, visualizing)

2. Compatibility with Google Colab, Python ecosystem, and SQLite will help you to integrate.
3. Scalability: Possibility to extend functionality or manage more massive data
4. Performance: Under restricted computing resources, dependability and speed of execution
5. Preference for free, changeable tools under open-source licencing
6. User Community & Documentation: Possibility of support, examples, and community involvement.

DEEP DIVE INTO SELECTED TOOL- METABASE

Designed as an open-source business intelligence (BI) tool, Metabase lets users generate dashboards, charts, and reports from their data without writing much code. It interacts with PostSQL, MySQL, and SQLite among other databases. Using interactive charts and filters, Metabase was applied in this project to translate processed healthcare research data into visual insights.

Features & Capabilities:

1. lets users create dynamic dashboards using several filters.
2. supports a no-code query builder with SQL-based searches.
3. provides several chart forms including tables, bar graphs, pie charts, and line graphs.
4. Dashboards may be included into other systems or sent via links.
5. Simple interface fit for nontechnical as well as technical people.

Target Clients:

Metabases is most appropriate for:

1. Small to medium companies in search of a free analytics tool.
2. Consultants and analysts handling several client data sources.
3. Academic researchers in need of fast data visualizations.
4. Teams seeking open-source substitutes for corporate BI systems.

Pricing & Licencing:

1. With an AGPL license, the open-source variant is free and offers all basic functionality.

2. Extra capabilities including role-based permissions, audit logs, and email alerts are provided by paid plans—Pro and Enterprise.

Technical Requirements & Integrations:

1. JAR files, Docker, or cloud deployment will enable Linux, Mac, or Windows running capability.
2. connects with widely used databases including SQLite, MySQL, Postgresql, MongoDB, and others.
3. Metabase was used locally under Docker in this project, linked to a SQLite database exported from Google Colab.
4. Using Docker on a Windows system, Metabase was configured. File path mapping first proved problematic, however hand configuration fixed these problems. Once linked, the tool operated flawlessly and dashboard development was quick and simple.

Pros:

1. One advantage is free, open-source.
2. Simple to operate and configure
3. Excellent help for filters and SQL searches.
4. interacts with several kinds of databases.
5. Ideal for interactive dashboards

Cons:

1. Comparatively to programs like Tableau, visuals offer less customizing.
2. Some sophisticated capabilities call for a premium version.
3. First Docker setup could call for troubleshooting.
4. Not perfect for reporting on a big scale without the Pro version

Comparative analysis using alternatives:

Tool	**Advantages**	**Drawbacks**
Metabase	Free, easy to use, compatible with SQLite	Limited visualization options, manual setup required

Tableau	Good community support, high-quality visualizations	Expensive, harder to learn
Power BI	Strong Microsoft integration, rich feature set	Requires a license, Windows-focused
Looker	Cloud-native, supports advanced data modeling	High cost, complex configuration

REAL-WORLD USE CASE

Business Scenario: Automation for Review of Cancer Research

For a university research team working on a grant for a cancer treatment, reviewing hundreds of scholarly papers from databases like PubMed is expected. With limited staff and tight deadlines, manually evaluating every paper is impractical, hence important findings could be lost. They want a quick method to spot trends in a lot of unorganized data and get important understanding.

How the Tool Solves the Problem:

From topic modeling to text extracting to summarizing, the AI-Based Healthcare Research Insights System does the whole process automatically. Important highlights from research PDFs might be produced via PyPDF2, spaCy, and a BERT-based summarizer allowing the system to Then latent dirichlet allocation (LDA) groups the papers into pertinent groups such as Cancer_Treatment_Usage. Metabase dashboards let users graphically investigate these insights by letting them interactively filter summaries, word clouds, and subject trends to rapidly concentrate relevant data.

Business Benefits:

The method cuts hand labor, saves time, and lets researchers make quick, data-based conclusions. the reasonable cost comes from the use of scalable for larger datasets or long-term use and open-source technologies. Interactive dashboards enhance usability even for non-technical users once put in use.

Risks and Considerations:

Among the restrictions it includes accuracy issues with small noisy datasets, need of basic technical skills to set up Docker and Metabase and the challenges with not sufficient metadata that may distort time based analysis. However crucial for ensuring continuing performance and compatibility is regular maintenance. Notwithstanding these difficulties our system offers effective and affordable way to streamline medical research analysis.

FINAL RECOMMENDATION

Summary of Judgment:

For fields where extensive amounts of scholarly material must be quickly evaluated, the AI-Based Healthcare material Insights System offers a solid, pragmatic option for automated literature reviews. Open-source, low-cost tools such PyPDF2, spaCy, BERT, LDA, and Metabase allow expertly coupled text extraction, summarizing, topic modeling, and interactive dashboards.. Drawing on our knowledge, the system provides useful insights with little manual effort and is both scalable and functional.

When to Recommend / When Not To:

For academic organizations, small-to- mid-sized research teams, and healthcare analysts that must rapidly and at scale process research publications, we suggest this approach. It is quite fitting when the goal is to compile summaries, identify issue clusters, and display results for the development of proposals or decisions. Its open-source nature qualifies it as a reasonably cheap replacement for systems of commercial literature analysis.

This one might be perfect for consumers without technical knowledge—especially those not familiar with Python, Docker, or SQL-based tools—even if this approach would not be ideal. It may also fail in extremely small or low-quality datasets where topic modeling loses dependability. Applications in clinical or regulatory environments involving high stakes would call for further validation procedures.

Final Thoughts on Client Fit:

This method is often appropriate as a consultation tool for clients looking for a scalable, reasonably priced, configurable approach to transform unstructured academic material into ordered, practical insights. Under appropriate technical direction and continuous assistance, it

may be a great resource in research-driven decision-making throughout healthcare, education, policy, and beyond.

APPENDIX

Figure A1: AI-Based BERT Summarization Output

Extractive summaries generated by the BERT-based model from input articles.

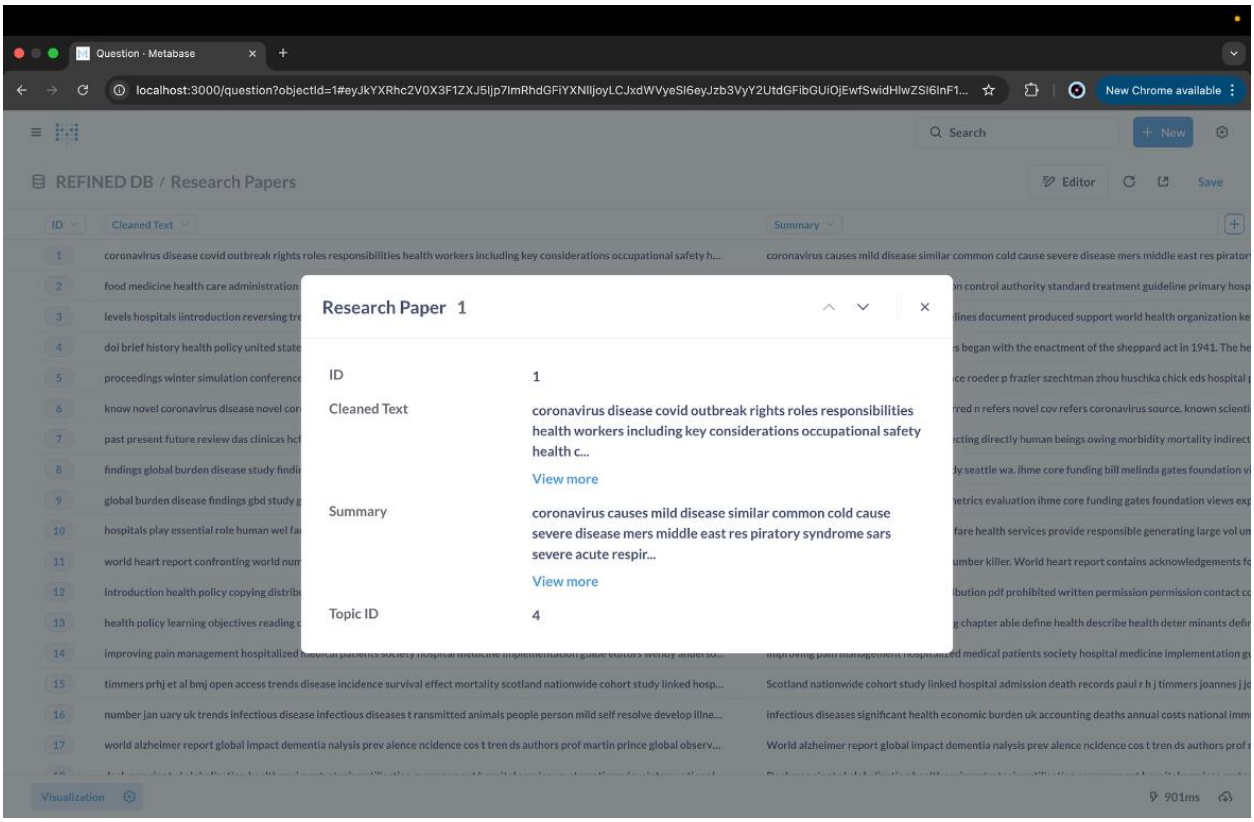


Figure A2: TOPIC MODELLING USING BERT

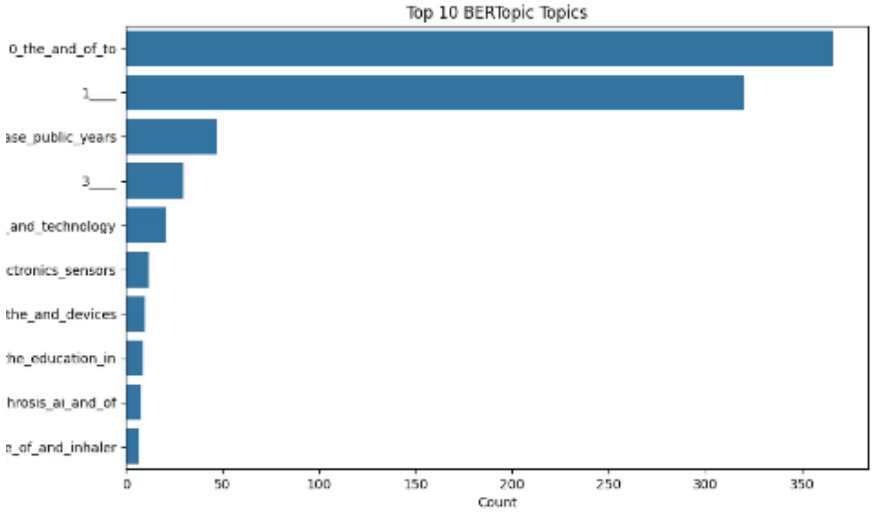


Figure A3: TOPIC MODELLING USING TRADITIONAL LDA

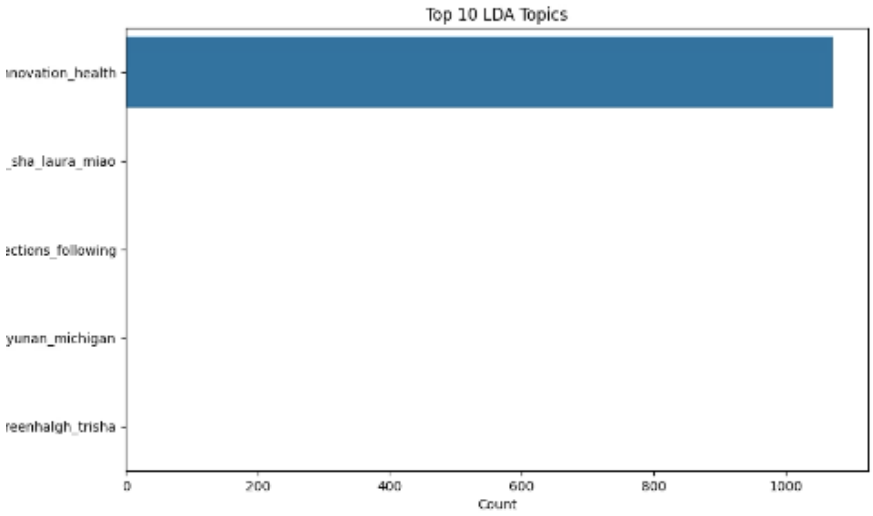
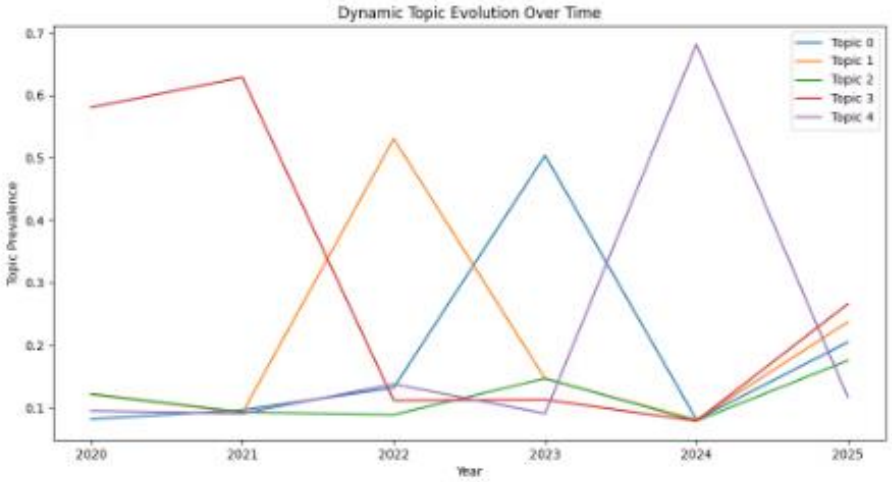


FIGURE A4: DYNAMIC TOPIC EVOLUTION OVER TIME



ADVANCED VISUALIZATIONS

FIGURE A5: HEAP MAP

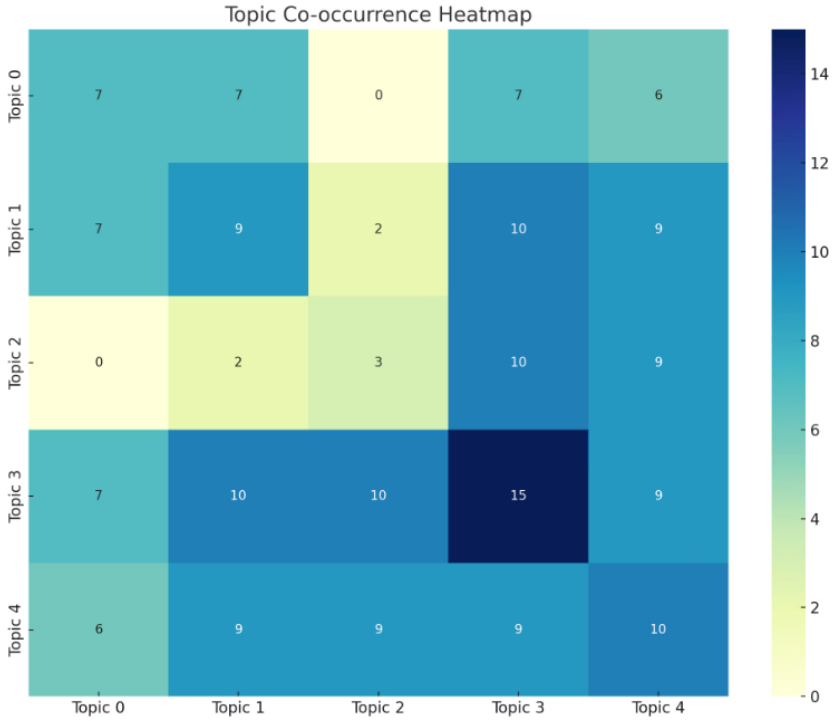


FIGURE A6: BUBBLE CHART

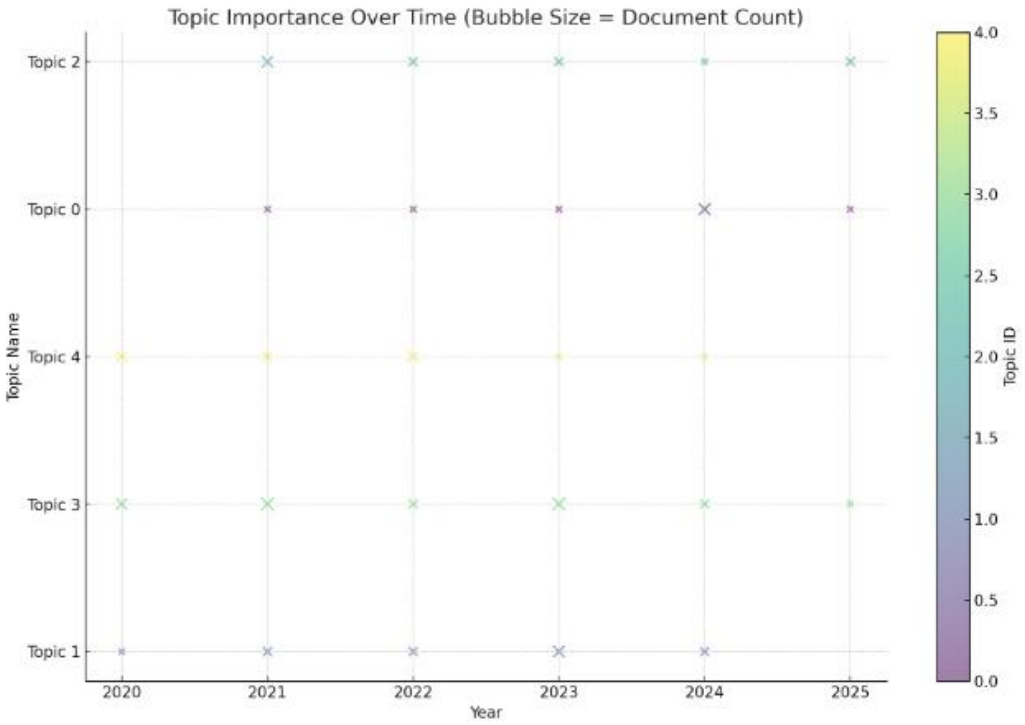


FIGURE A7: DASHBOARD VIEW

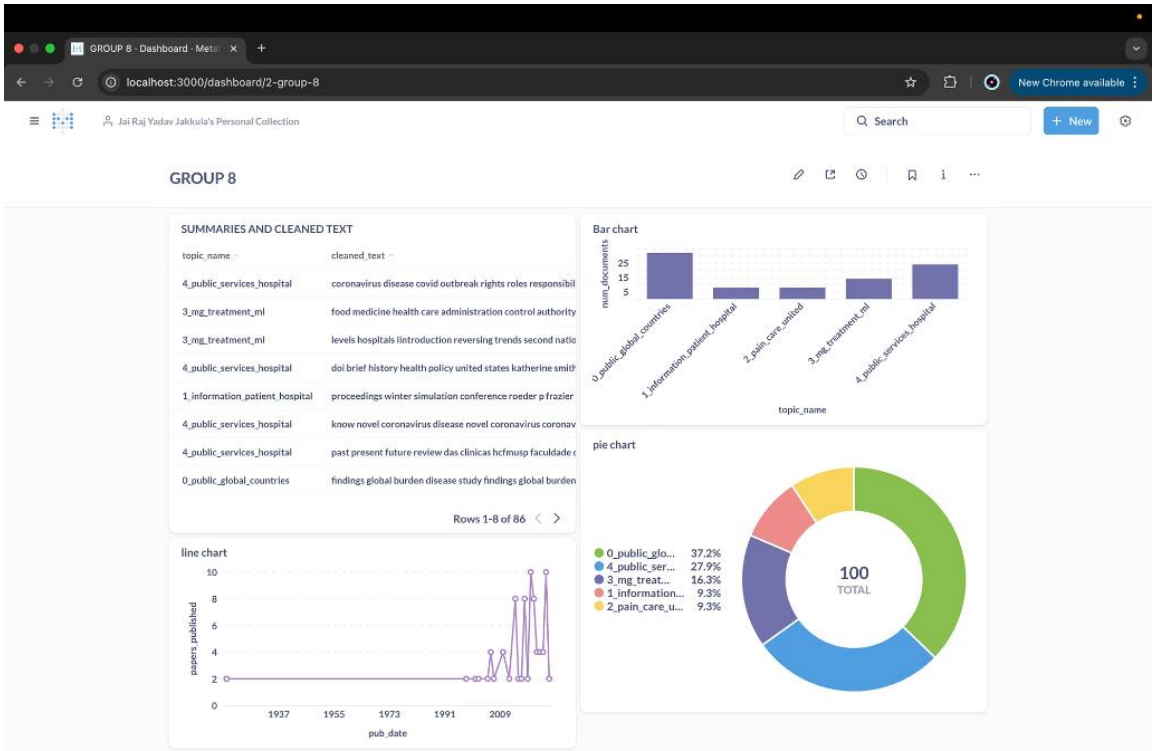


FIGURE A8: PIE CHART- TOPIC-WISE DISTRIBUTION OF RESEARCH PAPERS

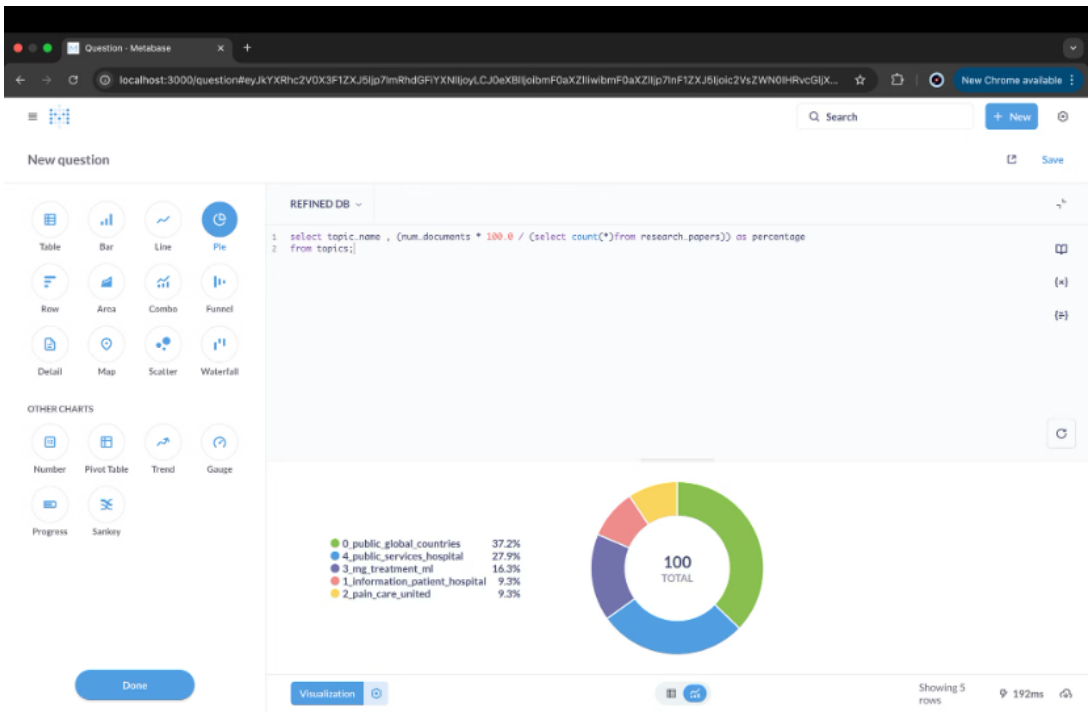


FIGURE A9: BAR CHART- TOP TOPICS BY NUMBER OF PAPERS

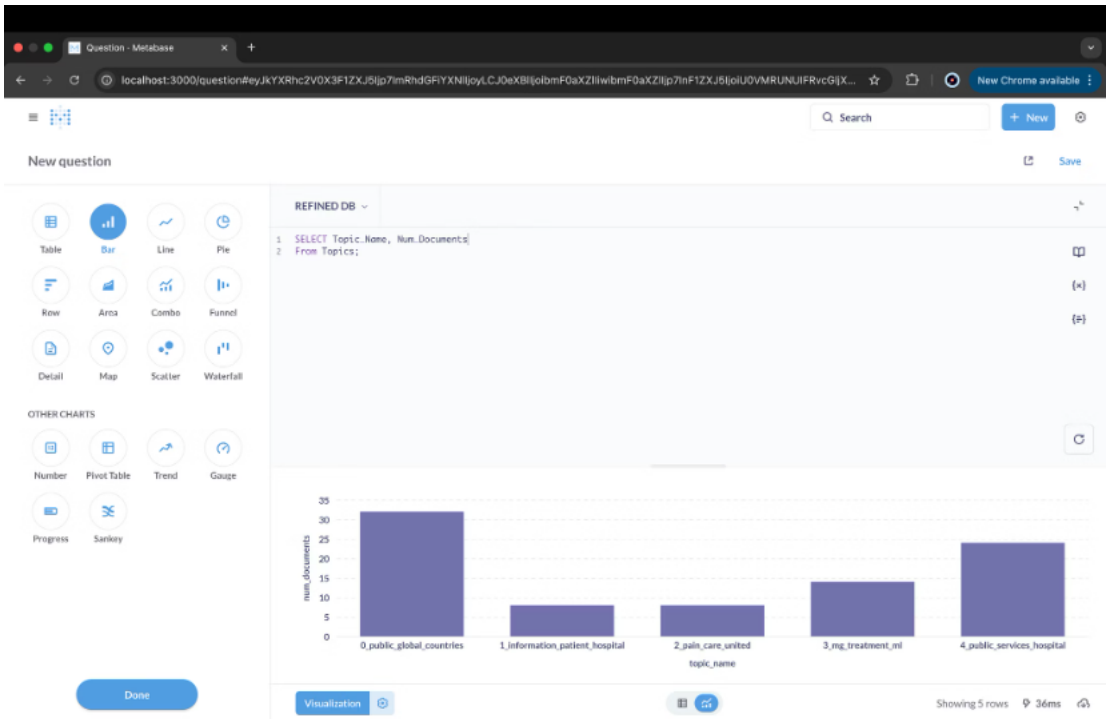


FIGURE A10: LINE CHART- RESEARCH PAPER TRENDS OVER TIME

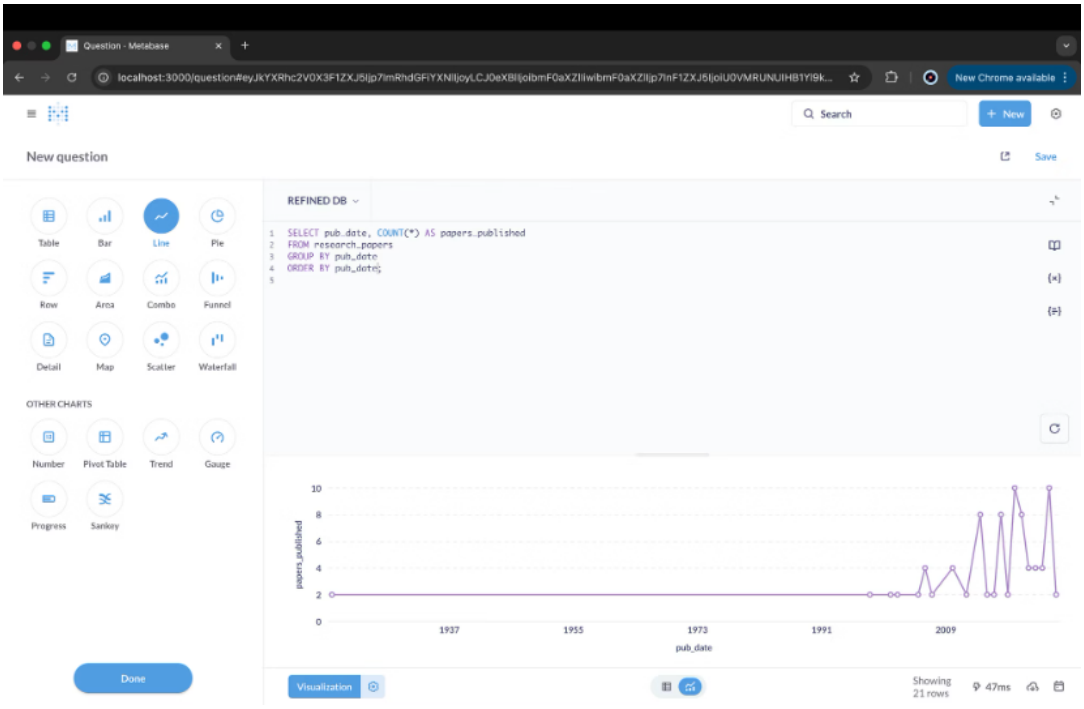
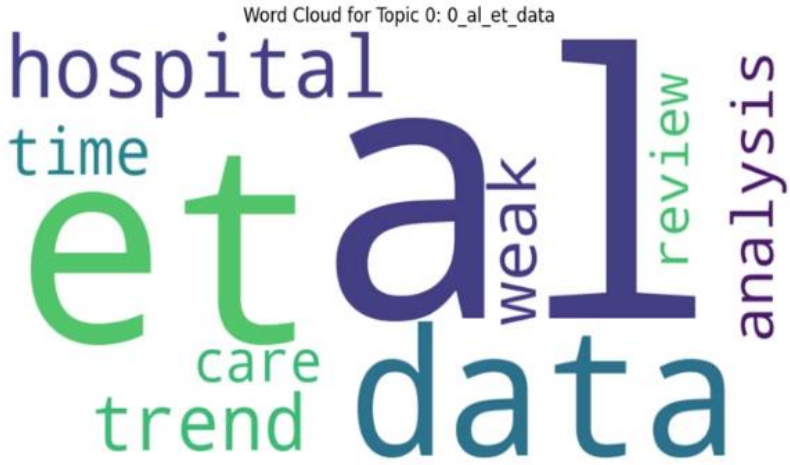
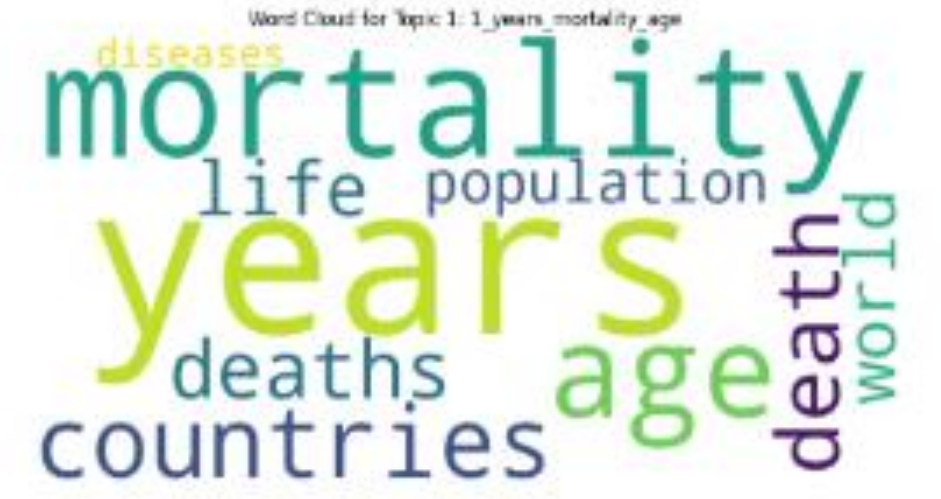


FIGURE A11: WORD CLOUD- MOST FREQUENT KEYWORDS ACROSS RESEARCH PAPERS

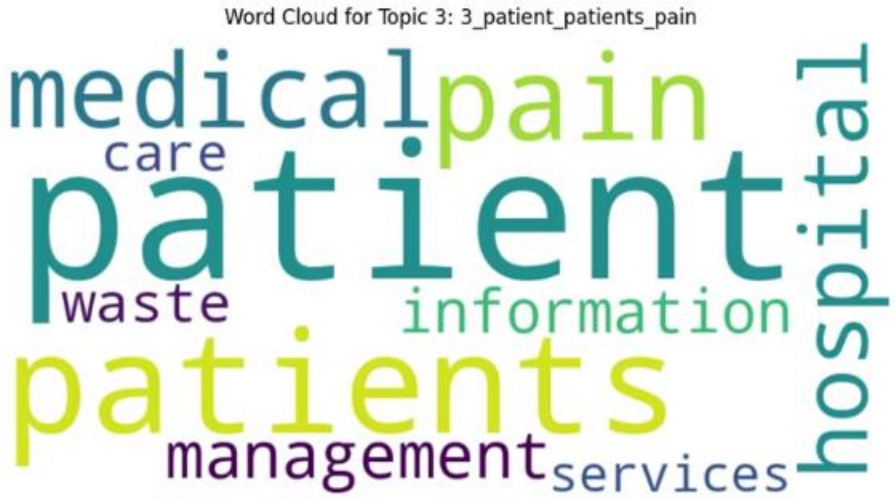
Topic 0:



Topic 1:



Topic 3:



Topic 4:


Word Cloud for Topic 4: 4_public_services_policy
programs
services
care
global
policy
social
public
states
government
federal

REFERENCES

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In Proceedings of NAACL-HLT. https://arxiv.org/abs/1810.04805

Grootendorst, M. (2022). *BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics*. https://maartengr.github.io/BERTopic/

Řehůřek, R., & Sojka, P. (2010). *Software Framework for Topic Modelling with Large Corpora*. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. http://radimrehurek.com/gensim/

Explosion AI. (2023). *spaCy: Industrial-strength Natural Language Processing in Python*. https://spacy.io

Python Software Foundation. (2023). *PyPDF2: PDF toolkit in Python*. https://pypi.org/project/PyPDF2/

Metabase. (2024). *Metabase Documentation*. https://www.metabase.com/docs/latest/

Docker Inc. (2024). *Docker Documentation: Get Started with Docker*. https://docs.docker.com/get-started/

SQLite Consortium. (2024). *SQLite Documentation*. https://www.sqlite.org/docs.html

BMC Public Health. (2021). *Perceived risk, anxiety, and behavioural responses of the general public during the early phase of the Influenza A (H1N1) pandemic in the Netherlands*. https://doi.org/10.1186/1471-2458-11-117