MINE-DD: Mining the Past to Protect Against Diarrheal Disease in the Future

Project Overview

The MINE-DD project (Mining the past to protect against diarrheal disease in the future) addresses the critical challenge of predicting and preparing for diarrheal disease burden in a climate-changing world. As global climate change manifests through altered precipitation patterns, flooding, and drought, it threatens to reverse decades of progress in reducing diarrheal disease—the second leading cause of death in children globally.

Diarrheal disease emerges from complex interactions between pathogens, human hosts, infrastructure, and environmental factors. Current disease surveillance and policy prioritization are limited by insufficient understanding of the underlying causal pathways connecting climate variables to disease outcomes.

Project Goals and Methodology

The project employs artificial intelligence to extract knowledge from existing scientific literature, using this information to project future diarrheal disease risk patterns. Specifically, MINE-DD applies semi-supervised Information Extraction techniques from Natural Language Processing to systematically analyze fifty years of scientific publications across different geographies and time periods.

The primary objective is to extract statistically significant associations between climate-sensitive variables and pathogen-specific diarrheal disease prevalence. This extracted information will be used to construct a system dynamics model capable of projecting pathogen-specific enteropathogen disease burden under various future climate scenarios.

Expected Outcomes and Impact

The resulting projections and insights will provide communities and policymakers with valuable tools to prepare for, adapt to, and build resilience against the inevitable health impacts of climate change on diarrheal disease. By enabling evidence-based decision-making, the project aims to guide policy formation and resource allocation in both high and low-income settings.

How Embeddings Work in the Project

Embeddings are a crucial component of the MINE-DD project's natural language processing workflow. Here's how they function within the project context:

1. **Text Representation**: Embeddings convert text (such as scientific papers about diarrheal diseases) into numerical vectors in a high-dimensional space. Each document or text segment is transformed into a series of numbers that capture its semantic meaning.

2. **Semantic Understanding**: In the PaperQA framework used by MINE-DD, embeddings (like "ollama/mxbai-embed-large") enable the system to understand the conceptual relationships between different documents. Similar concepts end up positioned closer together in the embedding space, even if they use different specific terminology.
3. **Efficient Knowledge Retrieval**: When a question is posed to the system (e.g., "How does temperature affect rotavirus incidence?"), the question is also converted to an embedding. The system can then efficiently find relevant scientific papers by computing the similarity between the question embedding and document embeddings.
4. **Cross-Document Connections**: Embeddings allow the system to make connections across documents, identifying subtle relationships between climate factors and specific disease outcomes that might not be explicitly stated in any single paper.
5. **Integration with Large Language Models**: The embeddings work in conjunction with large language models (like the Llama models used in MINE-DD) to first retrieve relevant information and then synthesize coherent answers from multiple sources.

This embedding-based approach enables the MINE-DD project to efficiently process vast amounts of scientific literature, extracting the specific climate-pathogen relationships needed to build accurate predictive models for future disease burden.