Data Science Take Home

Transforming Patient Transcripts into Structured Data for Medical Research Analysis

Executive Summary

The initiative focused on transforming unstructured patient transcript data into a structured SQL database format. By employing Named Entity Recognition (NER) technology, specifically the biomedical NER model trained on medical data, the project aimed to extract medically relevant entities. This approach enables efficient data analysis and supports the research objectives of the medical institute. The report delves into the methodologies employed, decision-making processes, and future enhancement suggestions.

Named Entity Recognition (NER) Process

Model Selection

The biomedical NER model was chosen for its specialized training on medical datasets, ensuring high accuracy in identifying relevant medical entities, thereby aligning closely with the project's requirements.

Methodology

After collecting the dataset, an initial Exploratory Data Analysis (EDA) was conducted to understand its structure. The dataset underwent basic preprocessing and cleaning to prepare it for the NER process, which identified 37 unique entities. These entities were then filtered based on their relevance and importance, considering the stakeholders' perspective.

Entity Selection Rationale

Selection was influenced by the entities' relevance to the institute's medical research goals, focusing on patient conditions, treatment outcomes, and diagnostic procedures. Important demographics such as age, sex, and race were retained for all user IDs, alongside critical features like medications, diseases, symptoms, and severity. Additional details such as diagnostic and therapeutic procedures, biological structures, and patient history, including family history, were considered essential for uncovering hidden patterns.

SQL Table Structure Design

To accommodate the diverse entities identified, strategies like nullable columns and a normalized design can be implemented to optimize data storage and handle sparsity. The decision-making process involved weighing the importance of each attribute, leading to a comprehensive SQL table structure that supports efficient data analysis and scalability.

Normalization and Entity Relationships

To accommodate the multivalued nature of some entities, the database schema can be normalized into separate tables, each focusing on a specific aspect of patient data. This approach not only reduces redundancy but also facilitates detailed analyses of medications, disease disorders, patient histories, and other significant factors without overwhelming the primary patient table.

- A **Patient** table encapsulates demographic details, acting as the primary entity

- **Medication**, **Disease_Disorder**, **History**, and additional tables for **Symptoms**, **Biological_Structure**, **Diagnostic_Procedure**, and **Therapeutic_Procedure** can be created as separate entities the multivalued entities, ensuring a clean, organized structure that mirrors the complexity of medical data.
- We can remove some entities based on the granularity of details required for the desired solution

Business and Technical Considerations

The project's technical solutions are intricately aligned with the medical institute's research objectives. The entity selection process significantly enhances analysis capabilities, fostering informed decision-making. Furthermore, the SQL design is optimized for cloud deployment, offering scalability, security, and ease of access. Utilizing cloud platforms such as AWS Redshift or Athena enables advanced analysis and visualization, e.g., through AWS Quicksight.

Recommendations and Future Work

- **NER Model Exploration**: Conduct further research to identify the optimal NER model for medical domains.
- **Entity Expansion**: Consider adding more entities in future iterations to enrich the dataset for a broader range of research questions.
- **SQL Schema Refinement**: As new data is ingested, the SQL schema may require adjustments to maintain alignment with evolving research needs.
- **Business Alignment Review**: Regularly review project outcomes to ensure they remain aligned with the institute's strategic objectives.
- **Human Feedback**: Incorporate human feedback to fine-tune the solution for improved accuracy and relevance.
- **Integration with LLM Models**: Leverage Large Language Models to provide user-friendly, quick, and insightful information.

Conclusion

This project marks a significant advancement in the use of patient transcript data for medical research. The application of NER technology and strategic SQL design has established a foundation for sophisticated data analysis, promising to significantly enhance the medical research institute's efforts with robust data-driven insights.