FOUNDATIONS OF MACHINE LEARNING -CS19643

Shaun Paul Moses 220701266
Someshwar K M 220701283

Introduction

This project aims to develop and evaluate methods for extracting and analyzing SDoH from clinical text using natural language processing (NLP) and machine learning techniques. By transforming unstructured data into actionable insights, the research seeks to support interventions that reduce health disparities and promote equity in healthcare delivery.

Problem Statement

Identifying Social Determinants of Health from Clinical Narratives Description: Propose methods to extract and analyze social determinants of health (SDoH) from unstructured clinical narratives. Research should aim to provide actionable insights that address health disparities and improve population health outcomes. give a small introduction for this project statement

Objective

- Identifying Social Determinants of Health from Clinical Narratives
- Suggesting remedies for identified social determinants of health and clinical narratives
- producing confidence score for identified SDoH labels
- Viusalizing processed data for clinical interpretation

Abstract

This project presents a **natural language processing (NLP)**-based framework to identify and analyze **Social Determinants of Health (SDoH)** from free-text clinical narratives. Leveraging a zero-shot classification pipeline using the **facebook/bart-large-mnli model**, the system **classifies** notes into specific SDoH categories—ranging from financial insecurity and food access to **psychosocial** issues like trauma, shame, and grief. Each identified factor is matched with evidence-based remedies such as therapy, community support programs, or medical interventions. The solution features a Gradio-based user interface supporting both single and bulk input (CSV), dynamic visualizations including heatmaps and histograms, and exports actionable reports. This approach empowers public health systems and providers to uncover hidden disparities, facilitate early intervention, and inform policy decisions aimed at reducing health inequities and improving population-level outcomes.

Existing Solutions

- **SDoH NLP Challenge by n2c2 (2019)**
- Organized by National NLP Clinical Challenges (n2c2).
- Focused on extracting SDoH concepts like housing, employment, and insurance from clinical texts.
- Provided annotated datasets for benchmarking.
- **EMERSE (Electronic Medical Record Search Engine)**
- Developed at the University of Michigan.
- Allows clinicians to search unstructured notes for SDoH-related keywords.
- Primarily keyword-based, limited contextual understanding.

Proposed Solution

1. Model Selection

Model Used: facebook/bart-large-mnli

Technique: Zero-shot classification — allows classification into pre-defined labels without requiring training on task-specific data.

Candidate Labels: A curated list of 20 SDoH-related issues (e.g., "financial insecurity", "depression", "trauma").

2. INPUT MODES

Single Input (Text Box) = for one of note analysis

Bulk Input (CSV) = For large-scale batch processing.

3. Processing Pipeline

3.1 For Single Input:

- **Text Classification:** Run the note through the zero-shot pipeline.
- **Top Label & Confidence:** Extract the highest-confidence SDoH label.
- **Remedy Retrieval:** Match label to a predefined remedy.
- **Heatmap Visualization:** Create a single-row heatmap showing the model's confidence for all categories.

3. Processing Pipeline

3.2 For Bulk Input:

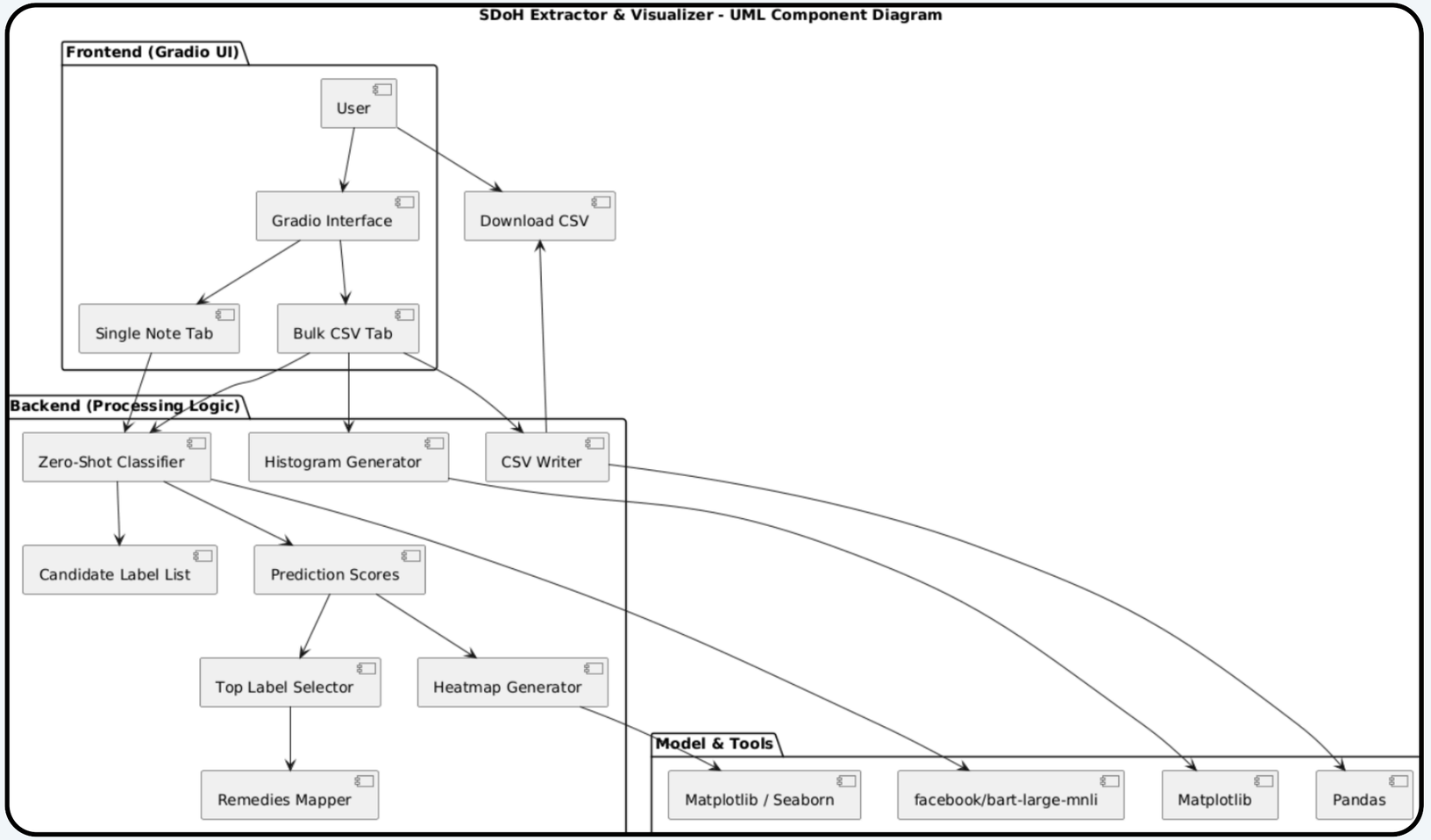
- **Read CSV:** Ensure it has a note column.
- **Parallel Classification:** Use ThreadPoolExecutor to process notes concurrently.
- **Per-note Output:**
  - Top SDoH label.
  - Confidence score.
  - Suggested remedy.

3. Processing Pipeline

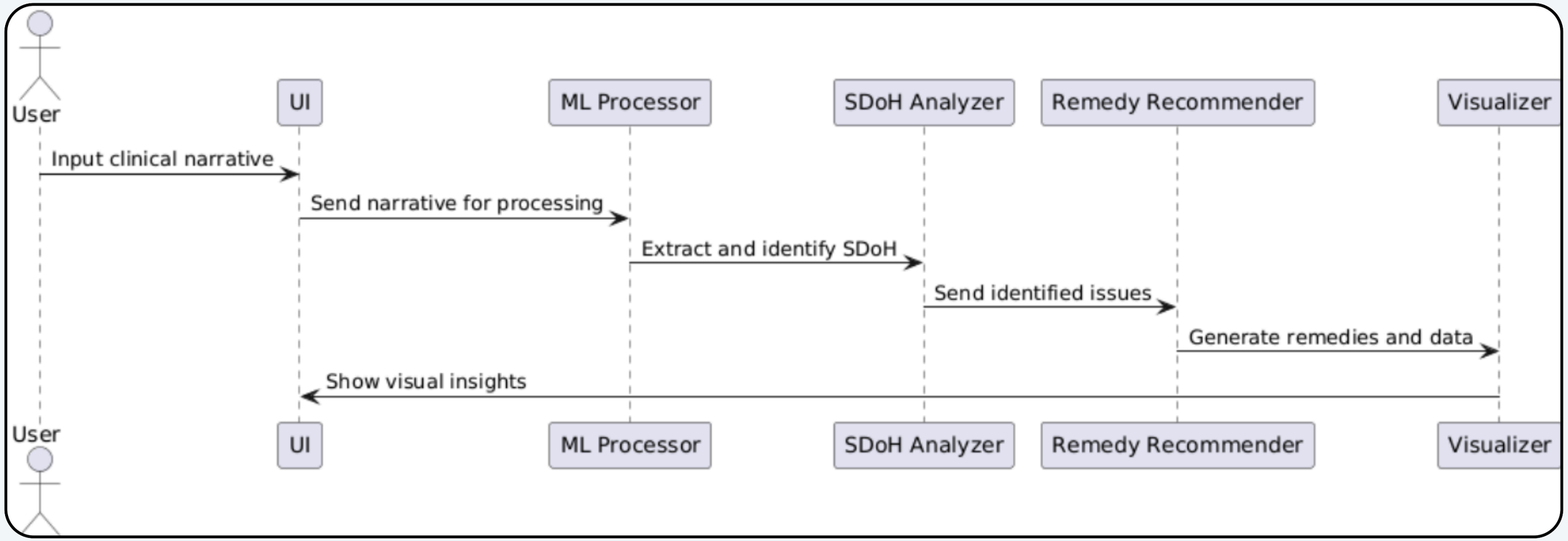
3.2 For Bulk Input:

- **Confidence Matrix:** Create a data frame of scores for all labels across notes.
- **Visualization:**
  - **Histogram:** Frequency of predicted SDoH labels.
  - **Heatmap:** Confidence scores across all notes and categories.
- **Output Files:** Export results and visualizations, provide CSV download.

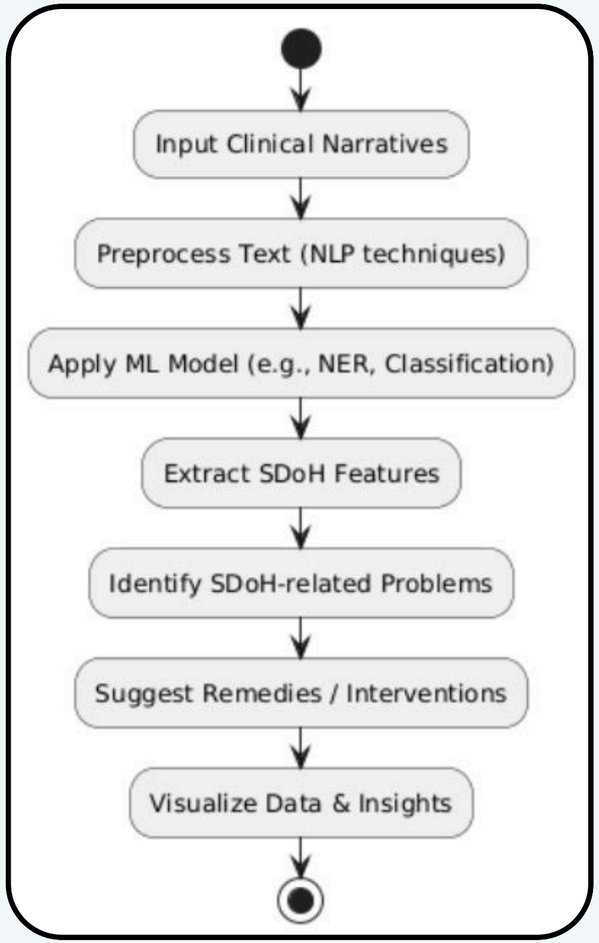
Architecture diagram



Sequnce Diagram



System Flow Diagram



Technologies Used

NLP Model : HuggingFace Transformers pipeline

GUI : Gradio (Blocks, Tabs, Buttons)

Visualization : Matplotlib, Seaborn

Data Handling : Pandas

Parallelism : ThreadPoolExecutor

Conclusion

This project demonstrates a practical and scalable approach to identifying Social Determinants of Health from unstructured clinical text using zero-shot learning and intuitive visualizations. By integrating NLP with interactive tools like Gradio, it not only extracts meaningful insights but also presents them in a user-friendly manner to support clinical decision-making and public health interventions.

References

[1] "Social Determinants of Health: The Solid Facts" - Marmot & Wilkinson Covers core SDoH concepts and evidence-based frameworks IEEE, 2024.

[2] Kumar, P., et al. **"Human Activity Recognitions in Handheld Devices Using Random Forest Algorithm."** In 2024 International Conference on Automation and Computation (AUTOCOM), pp. 159-163. IEEE, 2024.

References

[3] Kumar.P., et al. "Improvement of Classification Accuracy in ML Algorithm by Hyper-Parameter Optimization." In 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), pp. 1-5. IEEE, 2023.

[4] "Natural Language Processing with Python" - Bird, Klein, & Loper Foundational NLP techniques for text extraction (e.g., NER, tokenization). (2019)

THANK YOU