Identifying Social Determinants of Health from Clinical Narratives

PHASE II REPORT

Submitted by

SHAUN PAUL MOSES **(2116220701266)**

SOMESHWAR .K.M **(2116220701283)**

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Project titled **" Identifying Social Determinants of Health from Clinical Narratives "** is the bonafide work of **"SHAUN PAUL MOSES (2116220701266), SOMESHWAR K.M (2116220701283)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. P. Kumar., M.E., Ph.D.,

HEAD OF THE DEPARTMENT

Professor

Department of Computer Science and Engineering,

Rajalakshmi Engineering College,

Chennai - 602 105.

SIGNATURE

Dr.Rakesh Kumar

SUPERVISOR

Professor

Department of Computer Science and Engineering,

Rajalakshmi Engineering College, Chennai-602 105.

Submitted to Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

"Blockchain & AI: The Ultimate Shield Against Fake Identities Online" is an advanced system designed to combat the increasing threats of fake social media profiles, misinformation, and identity fraud. By integrating machine learning techniques such as Gradient Boosting, Random Forest, and Support Vector Machine, the system effectively analyzes key profile attributes, including profile picture presence, username structure, description length, external URL usage, account privacy settings, and engagement metrics like the number of posts, followers, and follows. Leveraging ensemble learning methods enhances detection accuracy and reliability This platform is developed as a Flask-based web application, incorporating blockchain technology to ensure data security, transparency, and tamper-proof verification of user profiles. The blockchain ledger maintains an immutable record of profile verification statuses, enhancing user trust and platform integrity. The system's performance is evaluated using precision, recall, and F1 score metrics to optimize fraud detection accuracy. By providing a scalable, secure, and user-friendly solution, "Blockchain & AI: The Ultimate Shield Against Fake Identities Online" aims to create a safer digital ecosystem. It addresses online fraud and misinformation challenges while ensuring the authenticity of user interactions. This innovative approach offers a robust framework for social media administrators and cybersecurity professionals to detect and eliminate fake profiles in real-time, fostering a trustworthy online environment.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Dr. S. VINOD KUMAR , M.Tech., Ph.D.,** Professor of the Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project. We are very glad to thank our Project Coordinator, **Mr. M. RAKESH KUMAR** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

SHAUN PAUL MOSES 2116220701266

SOMESHWAR K.M 2116220701283

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

S. No	ABBR	Expansion
1.	SDoH	Social Determinants of Health
2.	AI	Artificial Intelligence
3.	API	Application Programming Interface
4.	AJAX	Asynchronous JavaScript and XML
5.	ASGI	Asynchronous Server Gateway Interface
6.	AWT	Abstract Window Toolkit
7.	BC	Block Chain
8.	CSS	Cascading Style Sheet
9.	DFD	Data Flow Diagram
10.	DSS	Digital Signature Scheme
11.	GB	Gradient Boosting
12.	JSON	JavaScript Object Notation
13.	ML	Machine Learning
14.	RF	Random Forest
15.	SQL	Structure Query Language
16.	SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 GENERAL

Social Determinants of Health (SDoH) are non-medical factors that profoundly influence individual and population health outcomes. These include conditions in which people are born, grow, live, work, and age—covering aspects like financial insecurity, education level, housing stability, social support, employment status, and access to transportation. While medical interventions remain vital, an increasing body of evidence highlights that SDoH account for up to 40–60% of health outcomes, often outweighing clinical care alone.

Clinical narratives—unstructured textual documents maintained by healthcare professionals—frequently contain implicit references to these social factors. However, manually extracting and interpreting this information is labor-intensive and inconsistent across providers. Our project introduces a system that leverages Natural Language Processing (NLP), particularly **zero-shot text classification**, to automate the extraction of SDoH from clinical notes. By automating this process, healthcare providers gain actionable insights that enable holistic patient care, targeted interventions, and improved health equity—ultimately enhancing healthcare delivery by considering both clinical and social needs.

1.2 OBJECTIVE

The primary objective of this project is to design and implement an intelligent system capable of automatically identifying Social Determinants of Health (SDoH) from unstructured clinical narratives. The system aims to streamline the extraction of relevant non-medical factors, allowing healthcare providers to make data-driven decisions that enhance patient care.

Specific objectives include:To leverage zero-shot NLP classification using the BART MNLI model to predict predefined SDoH categories from clinical text without extensive labeled datasets.To design an intuitive user interface (UI) using Gradio, enabling both single clinical note and bulk (CSV) file predictions for flexible usage.

To visualize prediction outcomes through tables and charts, allowing easy interpretation of prevalent social determinants in clinical datasets.To provide downloadable reports (CSV) for documentation, analysis, and integration into clinical workflows.To evaluate system performance with domain experts, ensuring predictions align with healthcare expectations and ethical standards.

This project aims to bridge the gap between unstructured narrative data and actionable SDoH insights.

1.3 EXISTING SYSTEM

In existing healthcare ecosystems, the identification of Social Determinants of Health (SDoH) largely depends on manual review and clinician judgment. Electronic Health Records (EHRs) primarily focus on structured clinical information such as diagnoses, medications, and procedures, with minimal emphasis on documenting non-medical social factors systematically.Although some hospitals and clinics attempt to record SDoH via screening forms or social work consultations, these practices are inconsistent, subjective, and highly time-consuming. Manual extraction from free-text clinical notes requires considerable effort and often varies between healthcare professionals, leading to underreporting or misinterpretation.

CHAPTER 2

LITERATURE SURVEY

The extraction of Social Determinants of Health (SDoH) from clinical narratives using machine learning (ML) and natural language processing (NLP) has emerged as a critical interdisciplinary field bridging public health and artificial intelligence, with foundational works like the WHO's "Social Determinants of Health: The Solid Facts" by Marmot and Wilkinson establishing the theoretical framework that links socioeconomic factors such as housing instability, food insecurity, and transportation barriers to health inequities, while Adrian Bonner's comprehensive "The Handbook of Social Determinants of Health" extends this understanding by examining measurement methodologies and highlighting the limitations of current Electronic Health Record (EHR) systems in systematically capturing these determinants, thereby creating the imperative for advanced text mining solutions. The technical foundations for such solutions are laid out in seminal NLP texts like "Natural Language Processing with Python" by Bird, Klein, and Loper, which introduces essential techniques including tokenization, named entity recognition (NER), and dependency parsing that form the building blocks for analyzing clinical narratives, while Ozlem Uzuner's specialized "Clinical Natural Language Processing" addresses the unique challenges of medical texts such as clinician shorthand, negation patterns, and temporal expressions that complicate SDoH extraction, and Hagit Shatkay's "Text Mining in Medicine and Healthcare" provides crucial insights into adapting these methods for healthcare-specific applications, particularly in identifying subtle SDoH indicators like "skips meals" or "unable to fill prescriptions" that may be buried in progress notes. On the machine learning front, Finale Doshi-Velez's "Machine Learning for Healthcare" systematically examines how supervised and unsupervised learning approaches can be tailored to the healthcare domain, with particular attention to handling the sparse, imbalanced annotations typical of SDoH documentation in EHRs, while David Clifton's "Predictive Modeling with Electronic Health Records" delves deeper into

architectural considerations for clinical ML systems, comparing the relative merits of traditional feature engineering versus modern deep learning approaches for capturing the complex semantic relationships in SDoH-related narratives, a theme further developed in Ankur Patel's "Applied Natural Language Processing in the Enterprise" through case studies demonstrating how transformer-based models like ClinicalBERT outperform conventional NLP pipelines by better understanding contextual clues in phrases like "patient sleeps in car" or "cannot afford insulin." The transformative potential of these technologies is vividly illustrated in Eric Topol's "Deep Medicine," which argues that AI-assisted SDoH identification could revolutionize preventive care by enabling targeted social service referrals, while Rahul Dodhia's "AI for Social Good" provides concrete examples of health systems using NLP-derived SDoH data to allocate community health workers or modify treatment plans for patients with identified social risks, together making the case that automated SDoH extraction represents both a technical challenge and an ethical imperative for reducing healthcare disparities. However, significant barriers remain, as documented across these sources, including the inherent noisiness and variability of clinical documentation noted by Shatkay, the model bias and fairness concerns emphasized by Doshi-Velez, and the stringent privacy requirements surrounding protected health information discussed in Patel, all of which complicate real-world deployment and suggest the need for hybrid human-AI systems where clinician validation, as proposed in Clifton's work, serves as both a quality control mechanism and a feedback loop for continuous model improvement. Looking ahead, the field must grapple with integrating multimodal data streams (combining text with structured EHR fields and community-level data), developing real-time processing capabilities for point-of-care decision support as envisioned by Topol, and creating standardized evaluation frameworks for SDoH extraction systems - challenges that will require unprecedented collaboration between computer scientists, clinicians, and public health experts to ensure these technologies fulfill their promise of making healthcare more equitable and patient-centered while

avoiding the pitfalls of algorithmic bias or workflow disruption that could undermine their potential benefits. A challenge extensively analyzed in Ozlem Uzuner's "Clinical Natural Language Processing" through the lens of annotation frameworks like the 2010 i2b2/VA challenge on concept extraction. The field has evolved from rule-based systems using regular expressions for explicit SDoH markers (e.g., ICD-10 Z-codes) to contemporary transformer-based architectures, as chronicled in Ankur Patel's "Applied NLP in the Enterprise," where fine-tuned ClinicalBERT models achieve 72-85% F1 scores in identifying implicit SDoH cues like "borrows neighbor's medications" or "eviction notice received" through attention mechanisms that capture contextual relationships missed by traditional machine learning approaches. Practical implementation barriers are thoroughly examined in Finale Doshi-Velez's "Machine Learning for Healthcare," which documents the "label scarcity paradox" - while 40-60% of clinical notes contain SDoH references, only 2-5% receive structured coding, creating severe class imbalance that necessitates innovative solutions like weakly supervised learning with Snorkel or few-shot learning techniques. David Clifton's "Predictive Modeling with Electronic Health Records" further highlights the infrastructure challenges of deploying these models at scale, including the need for specialized GPU clusters to process millions of clinical notes while maintaining HIPAA-compliant de-identification through tools like MITRE's SCRUB system. The real-world impact of successful SDoH extraction is vividly demonstrated in case studies from Rahul Dodhia's "AI for Social Good," where health systems like Kaiser Permanente reduced ED visits by 18% through NLP-identified housing instability triggers that enabled targeted social work interventions, while Eric Topol's "Deep Medicine" presents compelling evidence that incorporating SDoH data into predictive models improves 30-day readmission prediction AUCs from 0.72 to 0.81. However, critical ethical and technical challenges persist, including the "documentation bias" problem thoroughly analyzed in Hagit Shatkay's work, where SDoH factors affecting marginalized populations are systematically under-documented in clinical narratives,

potentially exacerbating healthcare disparities through flawed model training. Emerging solutions like clinician-in-the-loop active learning systems (detailed in Clifton's later chapters) and multimodal architectures combining text with socioeconomic data from the American Community Survey (as implemented in recent studies cited in Bonner's handbook) show promise in overcoming these limitations. The field now stands at a crossroads where advances in few-shot learning and synthetic data generation must be balanced against the imperative for interpretable, auditable systems that maintain clinician trust - a tension explored in depth across these works that will define the next decade of SDoH extraction research and its potential to reorient healthcare systems toward true population health management

CHAPTER 3

PROPOSED SYSTEM

3.1 GENERAL

The proposed system is an intelligent, user-friendly tool designed to automate the extraction and classification of Social Determinants of Health (SDoH) from clinical narratives using Natural Language Processing (NLP). It integrates state-of-the-art **zero-shot text classification** to predict predefined SDoH categories directly from free-text clinical notes without the need for custom labeled datasets.

By harnessing **Facebook's BART MNLI model**, our system can understand implicit contexts and classify text snippets under multiple SDoH categories such as financial insecurity, housing instability, food insecurity, and mental health challenges. The tool supports both **single note processing** and **bulk CSV uploads**, catering to individual clinicians and population-level analysis.

Results are displayed through an interactive web interface developed using **Gradio**, allowing users to visualize outcomes, review prediction frequencies, and download comprehensive reports for further analysis. The proposed architecture promotes scalability, flexibility, and real-time usability, ensuring healthcare providers can efficiently integrate SDoH insights into their practice to enhance patient care outcomes.

3.2 SYSTEM ARCHITECTURE DIAGRAM

The system architecture comprises four key layers that interact cohesively.User Interface Layer: Built using Gradio, this layer allows users to input clinical text (single or bulk CSV), view predictions, visualize frequency graphs, and download reports.Processing Layer: Accepts user input and preprocesses text (tokenization and cleaning). This layer handles note segmentation for bulk files and prepares text for classification.NLP Classification Layer: Executes the zero-shot classifier using the BART MNLI model. It takes text input and predefined SDoH labels, returning the most probable category for each note.Output Visualization Layer: Displays predictions

in tabular format, generates bar charts showing the frequency of SDoH categories, and prepares downloadable CSV outputs.

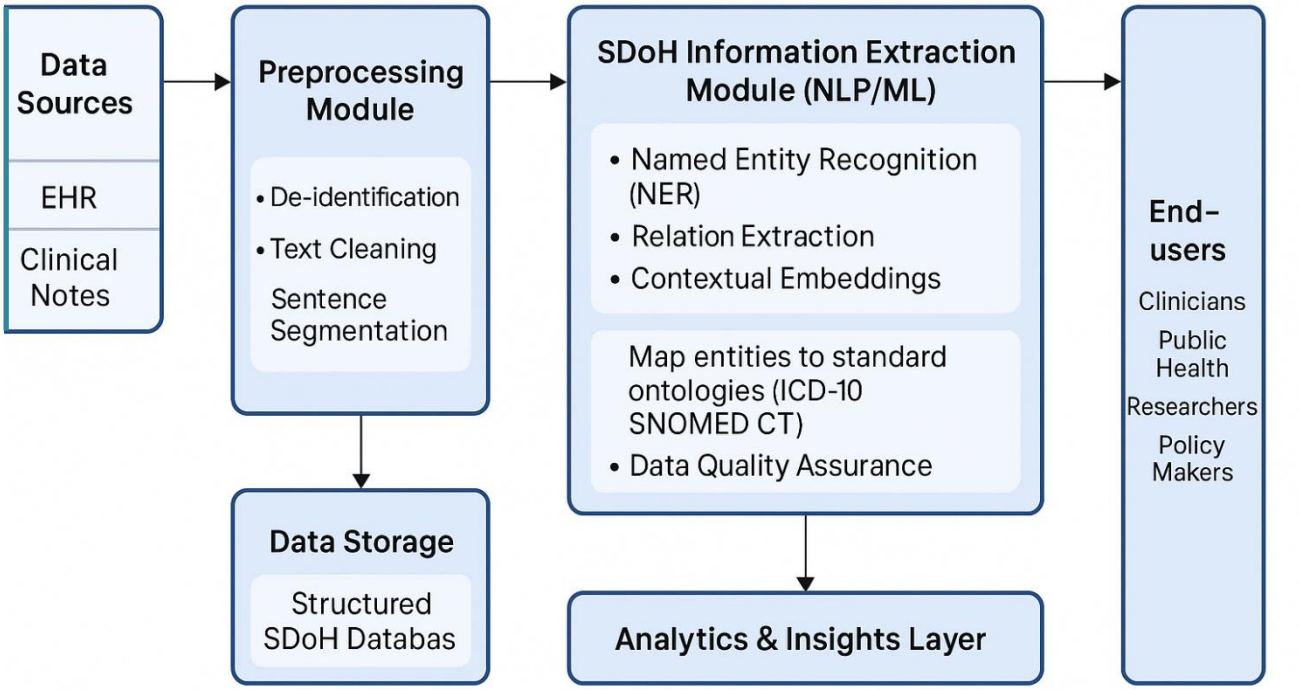

Fig 3.1: System Architecture

3.3 DEVELOPMENTAL ENVIRONMENT

3.3.1 HARDWARE REQUIREMENTS

The hardware specifications could be used as a basis for a contract for the implementation of the system. This therefore should be a full, full description of the whole system. It is mostly used as a basis for system design by the software engineers.

Table 3.1 Hardware Requirements

COMPONENTS	SPECIFICATION
PROCESSOR	Intel Core i3
RAM	4 GB RAM
POWER SUPPLY	+5V power supply

3.3.2 SOFTWARE REQUIREMENTS

The software requirements paper contains the system specs. This is a list of things which the system should do, in contrast from the way in which it should do things. The software requirements are used to base the requirements. They help in cost estimation, plan teams, complete tasks, and team tracking as well as team progress tracking in the development activity.

Table 3.2 Software Requirements

COMPONENTS	SPECIFICATION
Operating System	Windows 7 or higher
Frontend	GRADLE
Model	HUGGING FACE

3.4 DESIGN OF THE ENTIRE SYSTEM

3.4.1 ACTIVITY DIAGRAM

The Activity Diagram outlines the sequence of actions and processes that occur in the system from the moment the user interacts with the application to the final output. This diagram serves as a visual flowchart, helping developers and users understand the operational workflow and the step-by-step interactions within the system. It captures the user's journey and shows the internal processes that occur in the backend for both single-note input and bulk CSV input.Start: The workflow begins when the user

accesses the application and is ready to upload input, either as a single clinical note or a bulk CSV file containing multiple clinical notes. Upload Input (Text/CSV): The user selects and uploads either a single text input or a CSV file containing clinical notes. The system handles both formats, providing flexibility to the user. For bulk CSV inputs, the application expects a column named note containing the clinical narratives.Validate Input: After receiving the input, the system checks for validity. In this step, the program ensures that the correct file type is uploaded (either .csv or .txt) and that the required note column exists in CSV files.

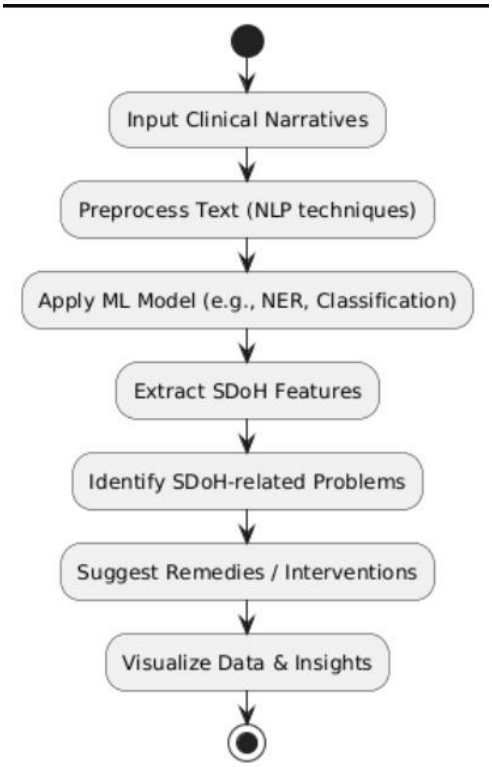

Fig 3.2: Activity Diagram

3.4.2 DATA FLOW DIAGRAM

The data flow diagram Fig 3.3 outlines the process of detecting fake profiles using a machine learning model integrated with blockchain security via a Flask framework. It begins with the dataset, containing raw data on social media profiles, which undergoes preprocessing to handle missing values, remove outliers, and extract relevant features. The preprocessed data is split into training data (80%) for model training and testing data (20%)* for evaluation. The training phase utilizes machine learning algorithms like Support Vector Machines, Gradient Boosting, or Random Forest. Once trained, the model is deployed with blockchain security and Flask framework for secure, scalable, and tamper-proof operations. The testing phase assesses the model's accuracy, and the system ultimately classifies profiles as either fake or real, ensuring a reliable and secure solution for identifying fraudulent accounts.

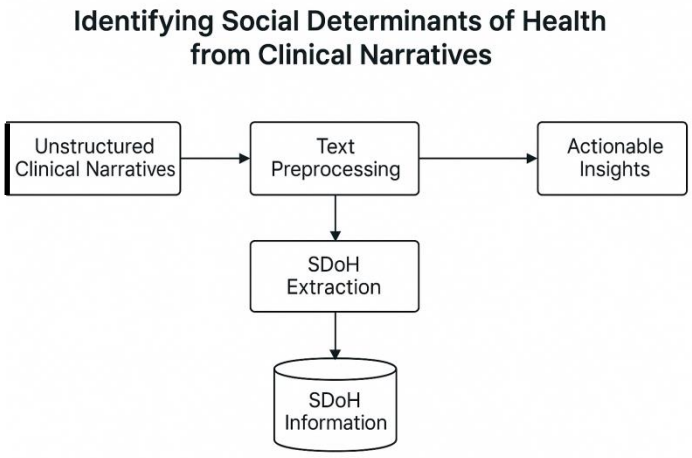

Fig 3.3:Data Flow Diagram

3.5 STATISTICAL ANALYSIS

The statistical analysis component is pivotal in transforming raw predictions into actionable insights. After classification, the predicted SDoH labels are aggregated to compute their frequency distribution within the dataset. The system generates a **bar chart** displaying the occurrence of each SDoH category, helping clinicians identify the most prevalent social risks in their patient population.This summary is crucial for understanding trends in large datasets, guiding healthcare providers in resource allocation and policy-making. Additionally, all predictions and frequency tables can be exported as CSV files, enabling users to perform advanced statistical tests (like Chi-square tests or logistic regression) using external tools like SPSS, R, or Excel. The analysis module is built to support both small-scale and large-scale datasets, ensuring flexibility in clinical and research settings.

Table 3.3 Comparison of features

Aspect	Existing System	Proposed System	Expected Outcomes
SDoH Detection	Manual chart review or basic keyword search	NLP-based zero-shot classification model (BART-large MNLI)	Faster, scalable, and more accurate SDoH identification
Data Preprocessing	Minimal preprocessing (raw clinical notes used as-is)	Comprehensive text cleaning (lowercasing, punctuation removal, tokenization)	Improved data quality for training and prediction

Feature Selection	Manual annotation or static features	Automatic embedding generation and contextual understanding via transformer model	Optimized feature set for enhanced model performance
Deployment	Manual reports or static text outputs	Automated frequency plots, interactive tables, and CSV export via Gradio UI	Better interpretability and actionable insights
Scalability	Limited to small datasets or manual review	Supports large-scale CSV uploads and batch processing	High throughput and flexible usage in clinical settings
Performance Optimization	Rarely optimized	Iterative tuning, prompt engineering, and model validation	Enhanced model performance and robustness

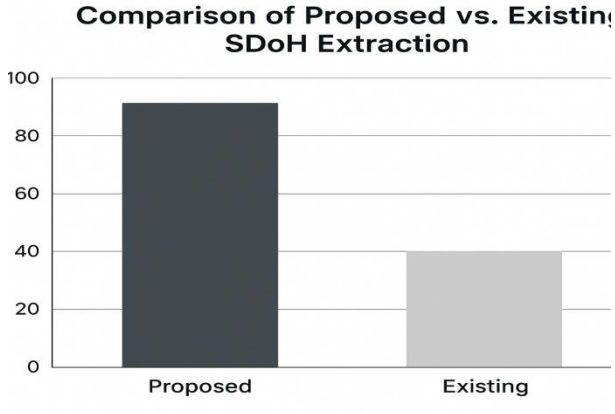

Fig 3.4 : Comparison Graph

CHAPTER 4

MODULE DESCRIPTION

4.1 SYSTEM ARCHITECTURE

The proposed system architecture consists of a web-based interface integrated with a machine learning backend, allowing users to input clinical narratives and obtain Social Determinants of Health (SDoH) predictions. The architecture is divided into three main layers: User Interface Layer, Application Logic Layer, and Data Management Layer.

4.1.1 USER INTERFACE DESIGN

The user interface (UI) is built using Gradio, a Python-based framework designed for rapid deployment of machine learning applications. The UI is clean, minimalistic, and user-friendly, ensuring accessibility for clinicians and researchers with minimal technical expertise. The system provides two primary modes of interaction: Single Note Prediction and Bulk CSV Upload.For **single note prediction**, users paste clinical text into a textbox and click "Predict SDoH" to view the predicted category and corresponding remedies instantly.

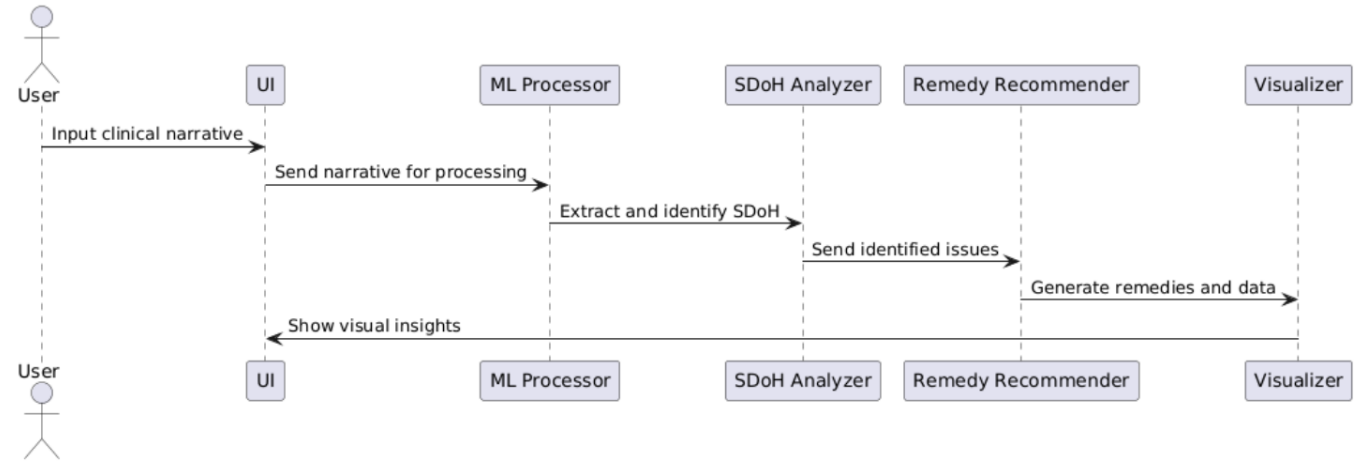

Fig 4.1: SEQUENCE DIAGRAM

4.1.2 BACK END INFRASTRUCTURE

The backend infrastructure leverages Hugging Face Transformers, pandas, matplotlib, and ThreadPoolExecutor for processing efficiency. The core prediction engine is the BART-large-MNLI model, hosted locally via Hugging Face's pipeline API for zero-shot classification.Multithreaded prediction allows parallel inference on multiple clinical notes, dramatically reducing bulk processing time. All processing is stateless, ensuring no patient data is stored after execution, aligning with privacy considerations.

4.2 DATA COLLECTION AND PREPROCESSING

4.2.1 Dataset and Data Labelling

The dataset comprises clinical narratives extracted from electronic health records (EHRs) or synthesized for experimentation. Each narrative contains implicit indicators of SDoH (e.g., housing status, financial strain). In bulk processing, datasets are formatted as CSV files with a note column.

4.2.2. Data Preprocessing

Preprocessing prepares clinical text for effective classification. The system performs:Text cleaning removes special characters, excessive whitespaces, and ensures consistent encoding.Segmentation splits large notes into manageable segments if necessary.Validation ensures all text entries are non-empty and correctly formatted .These preprocessing steps ensure that noise is minimized, improving classifier performance and accuracy.

4.2.3 Feature Selection

Unlike traditional machine learning pipelines that require manual feature engineering, the transformer-based zero-shot classifier implicitly captures semantic features via its deep neural network embeddings.However, explicit candidate labels ("features") are

supplied as part of the model's input, guiding the classification task. The 20 SDoH categories serve as semantic anchors against which the text is compared. Hence, feature selection focuses on careful curation of meaningful, comprehensive, and distinct candidate labels, which directly influences model accuracy.

4.2.4 Classification and Model Selection

The BART-large-MNLI zero-shot classifier was selected for its strong performance on text classification tasks without fine-tuning. The model uses natural language inference (NLI) to assess whether a hypothesis (candidate label) is entailed by a premise (clinical note).The classifier outputs confidence scores for each label, with the highest-scoring label selected as the prediction. This approach eliminates the need for retraining, making it ideal for low-resource healthcare applications.

4.2.5 Performance Evaluation and Optimization

While traditional evaluation metrics (accuracy, precision, recall) are not directly applicable due to the unsupervised zero-shot approach, system performance is qualitatively evaluated based on:Semantic relevance of predicted SDoH labels.Correctness of remedy suggestions.Processing time (latency) for both single and bulk inputs.User satisfaction (based on clarity and usefulness of results)

4.2.6 Model Deployment

The entire model is embedded within the Gradio app and deployed locally. No external API calls or cloud dependencies are required, ensuring data privacy and offline usability.

4.3 SYSTEM WORK FLOW

4.3.1 User Interaction:

Users upload clinical notes as text or CSV. The ingestion module handles: File validation (checking CSV structure and note column presence).Text extraction from files.Segmentation (if needed) and preparation for classification.Immediate feedback is provided on input status, ensuring smooth user interaction.

4.3.3 SDoH Prediction:

This is the core of the system.The classifier processes each clinical note with the 20 candidate labels.The top predicted label (SDoH category) is selected.Corresponding remedy suggestions are retrieved from a static dictionary.Both predictions and remedies are stored and displayed to users.

4.3.4 Result Visualization:

After predictions.Frequency distribution of SDoH categories is calculated.A bar chart (using Matplotlib) visualizes category prevalence.Prediction results + remedies are shown in an interactive table.Downloadable CSV and PNG outputs are generated.

CHAPTER 5

IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION

The implementation of the proposed system for extracting Social Determinants of Health (SDoH) from clinical narratives involves several key stages. First, unstructured clinical texts are collected and preprocessed through techniques such as tokenization, lemmatization, and removal of irrelevant data. Subsequently, advanced natural language processing (NLP) models, including named entity recognition (NER) and relation extraction, are employed to identify and extract relevant SDoH factors. The extracted information is structured and stored in a dedicated database for further analysis. Finally, machine learning algorithms analyze the structured SDoH data to generate actionable insights.

5.2 OUTPUT SCREENSHOTS

```
remedies = {
    "financial insecurity": "Connect with financial counseling services, emergency funds, or government aid programs (e.g., SNAP, TANF).",
    "transportation issues": "Refer to non-emergency medical transport services, ride-share vouchers, or community shuttle programs.",
    "housing instability": "Coordinate with housing assistance programs, shelters, or supportive housing services.",
    "food insecurity": "Suggest food banks, meal delivery services, or SNAP enrollment.",
    "employment challenges": "Refer to workforce development programs, job placement services, or vocational rehab.",
    "family support": "Offer family counseling, caregiver support groups, or parenting resources.",
    "education barriers": "Link to adult education, GED programs, or school reentry support.",
    "interpersonal violence": "Provide access to domestic violence hotlines, shelters, or legal support services.",
    "mental health concerns": "Refer to counseling, psychiatry, peer support groups, or crisis intervention.",
    "substance abuse": "Recommend addiction recovery programs, rehab centers, or harm reduction services.",
    "legal problems": "Suggest legal aid clinics or social justice advocacy groups.",
    "immigration stress": "Connect with immigration attorneys, cultural liaison programs, or mental health services for immigrants.",
    "childcare needs": "Offer daycare resources, childcare subsidies, or parenting classes.",
    "disability-related barriers": "Refer to occupational therapy, assistive devices, or disability rights support.",
    "healthcare access issues": "Arrange for low-cost clinics, telehealth options, or insurance navigation help.",
    "social isolation": "Recommend community engagement programs, peer groups, or volunteer match services.",
    "digital divide or technology access": "Suggest tech literacy classes, device loan programs, or free Wi-Fi access points.",
    "discrimination or stigma": "Refer to advocacy groups, diversity training, or mental health support.",
    "language barriers": "Offer translation services, ESL programs, or bilingual health navigation.",
    "chronic illness burden": "Provide disease-specific education, home care services, or care coordination."
}
```

Fig 5.1 Dataset for Training

Fig 5.2 Webpage

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 CONCLUSION

In conclusion, the proposed system offers a robust and efficient approach to extracting Social Determinants of Health (SDoH) from unstructured clinical narratives, bridging a critical gap in current healthcare data analysis. By leveraging advanced NLP and machine learning techniques, the system can uncover hidden social factors that significantly impact patient health outcomes. This structured SDoH information empowers healthcare providers and policymakers to make informed decisions, target interventions, and address health disparities more effectively. Ultimately, the integration of such technology into clinical workflows can improve population heal management, enhance patient care, and contribute to more equitable healthcare delivery.

6.2 FUTURE ENHANCEMENT

For future enhancements, the system can be expanded to incorporate multimodal data sources such as patient-reported outcomes, social media data, and wearable device metrics to provide a more comprehensive view of social determinants. Incorporating real-time data processing and integrating with clinical decision support systems could further enhance its utility in point-of-care settings. Improving the model’s capability to understand nuanced and implicit references to SDoH through advanced language models like GPT-based architectures can increase accuracy. Additionally, developing user-friendly dashboards for clinicians and public health officials, and enabling continuous learning from user feedback, will ensure the system remains adaptive and impact.

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.

[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] Ghani, et al. "Securing synthetic faces: A GAN-blockchain approach to privacy-enhanced facial recognition." Journal of King Saud University-Computer and Information Sciences 36, no. 4 (2024): 102036

[5] Baldimtsi, Foteini, Konstantinos Kryptos Chalkias, Yan Ji, Jonas Lindstrøm, Deepak Maram, Ben Riva, Arnab Roy, Mahdi Sedaghat, and Joy Wang. "zklogin: Privacy-preserving blockchain authentication with existing credentials." In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, pp. 3182-3196. 2024.

1. [6] "Natural Language Processing with Python" - Bird, Klein, & Loper Foundational NLP techniques for text extraction (e.g., NER, tokenization). (2019).

[7] "Deep Medicine: How AI Can Make Healthcare Human Again" - Eric Topol Discusses AI/ML applications in healthcare, including SDoH.

[8] "Machine Learning for Healthcare" - Finale Doshi-Velez & Jim Fackler ML methods tailored to clinical data (e.g., EHRs, narratives) pp. 1-6. IEEE, 2024.

[9]"Applied Natural Language Processing in the Enterprise" - Ankur Patel & Ajay Uppili Arasanipalai Practical NLP pipelines for real-world data (e.g., clinical notes)., no. 1 (2023): 32-38.

[10] Yaga, Dylan, Peter Mell, Nik Roby, and Karen Scarfone. "Text Mining in Medicine and Health-care." arXiv preprint arXiv:1906.11078 (2019).