Natural Language Processing Course Project - Initial Pitch

TITLE : MediQ - NLP based Medical Report Information Extraction & Visualization

Project GOAL: The aim is to simplify the interpretation of complex medical reports and integrate machine learning insights that can assist healthcare professionals in identifying potential conditions more efficiently.

Project Team Members

Team Size: 1

1. **[Sasidhar Gadepalli]** - [gadepalli.sa@northeastern.edu] - [Section 21600/21601]

Project Description

This project focuses on building an NLP-driven medical information system that extracts structured data from medical reports. Using techniques like entity recognition and value parsing, the system identifies key patient information such as demographics, lab results, diagnoses, and medications. The extracted data is then presented in a user-friendly dashboard, enabling clear visualization of health parameters with highlighted abnormalities. Beyond data visualization, the system also provides diagnostic suggestions based on abnormal lab values and clinical patterns, acting as a decision-support tool for healthcare professionals.(This part is still yet to be decided as to which approach to take). By automating the interpretation of complex medical documents, this project aims to reduce manual effort, improve accuracy, and enhance patient care.

Assumptions:

Due to the purpose of this course and the fact that getting real-world medical reports is hard and most of them are in the form of image data, this project uses artificial text-data, generated with the help of a custom python script. This is to ensure that the focus of the project remains on NLP aspects rather than making this a computer vision project involving OCR and image recognition. A sample medical report image was taken from Kaggle and the custom script was generated in such a way to generate text report data that closely resembles real medical reports. This ensures that our focus remains on the NLP aspects of the project.

Approach :

1. **Preprocessing Text reports :** Sentence Segmentation, Tokenization, Stop word removal, Lemmatization, Normalization, Regex Cleaning.
2. **Info Extraction :** NER, Value and unit extractions with Regex Parsing
3. **Storing in Structured format :** Storing the extracted neat info in json format to easily create dashboards.
4. **Dashboard Creation** : Using streamlit to build an appealing dashboard.
5. **ML based Diagnosis** : Based on the report suggesting, what the patient might be experiencing and what it may be using classification algorithms like Logistic regression, Random Forest or maybe Neural Networks.(the exact approach on this is yet to be finalized).

Dataset Information:

As mentioned before, since getting access to real world medical reports is complex, we used a custom script to generate our medical reports. However, these generated reports were based on a medical report that we used for reference from Kaggle (Image Reference). The python script generates reports such that our text reports are similar to what is being depicted in this report.

This helps us to create consistent and uniform reports that can be passed easily to the NLP tasks.

The dataset is structured to closely resemble real-world medical reports and contains several important fields. Each record begins with hospital information, followed by patient details such as name, ID, age, and gender, along with the consulting doctor's name and the date of the report. The core of the report is the laboratory results, which list tests like haemoglobin, RBC, WBC, platelets, and differential counts, along with their measured values, reference ranges, and abnormality flags (High/Low). In addition, the reports include clinical notes and diagnosis, which summarize the patient's medical conditions, as well as medications prescribed with details on drug name, dosage, and frequency. To make the data more realistic, the reports also contain noise or filler sentences, such as system verification notes or general doctor advice. These simulate the unstructured nature of real-world medical data and make the task more challenging.

For the annotation part, since the project uses synthetic data that already contains structured ground truth, a set reports will be converted into unstructured text and manually annotated for evaluation. The annotation plan covers patient details, lab tests with values and units, diagnoses, medications, and noise sentences, ensuring the model learns to capture only relevant entities. Model performance will be measured using precision, recall, and F1-score.

Sample data

```
Hospital: Smith, Flores and Hill Hospital
Patient: Joseph Watson, ID HSP20748, Age 51, Gender Male
Consulting Doctor: Dr. Kim Bradley, Date: 2025-10-02
Doctor advised proper rest and hydration.
Haemoglobin (g/dL) came out to be 15.19 g/dL, compared to normal
12.0-16.0.
Observed Total RBC (mill/cmm): 3.09 mill/cmm. Marked as L.
Haematocrit (%) was measured at 47.59 %. Marked as H.
WBC (/uL) was measured at 5944.62 /uL.
Lab recorded Platelets (/uL) value of 488695.52/uL. Marked as H.
Neutrophils (%) came out to be 73.28 %, compared to normal 40-70.
Marked as H.
Lymphocytes (%) came out to be 41.93 %, compared to normal 20-40.
Marked as H.
Observed Monocytes (%): 10.69 %. Marked as H.
Lab recorded Eosinophils (%) value of 3.14%.
Report verified by automated system.
Final Clinical Notes:
Diagnosis includes: Leukocytosis (Infection), Coronary Artery Disease
Medications prescribed:
 - Aspirin 81 mg, OD
 - Amoxicillin-Clavulanate 625 mg, TDS
Report verified by automated system.
```