NLP CA-1

NER in HealthCare

1. Introduction

Named Entity Recognition (NER) is a fundamental **Natural Language Processing (NLP)** task that involves identifying and labeling specific entities in text. In the **healthcare domain**, NER plays a vital role in extracting meaningful information, such as **medical conditions, treatments, medications, and symptoms**, from unstructured clinical or textual data.

This project focuses on extracting **medical entities** from healthcare-related text data. The goal is to automatically identify and categorize entities like **diseases, drugs, and symptoms**, which can streamline **clinical documentation, facilitate knowledge extraction, and support medical research**.

For this project, a **SpaCy-based NER model** was used due to its efficiency and pre-trained language representations. The model was fine-tuned on a **healthcare-specific dataset** containing medical text samples, with labeled entities categorized as **MEDICINE**, **MEDICALCONDITION**, and other relevant tags. This enables the model to recognize complex medical terms and accurately label them in unseen data, making it applicable for **healthcare text analysis and clinical decision support**.

2. Methodology

2.1 Data Preprocessing and Exploration

- The project uses a **healthcare-specific JSON dataset** containing text samples with labeled medical entities.
- Each sample consists of:
    - **Text content:** Describes medical cases, medications, or treatments.
    - **Annotations:** Labeled entities with `start`, `end`, and `tag_name` attributes.
- The dataset was processed to extract relevant fields:
    - **Text content** was extracted for NER processing.
    - **Annotations** were converted into a **SpaCy-compatible format** with entity positions and labels.
- The following NER categories were used:
    1. **MEDICINE** – Pharmaceutical drugs and treatments (e.g., Pepto-Bismol, loperamide).
    2. **MEDICALCONDITION** – Diseases, symptoms, and conditions (e.g., diarrhea, constipation).
- The dataset was tokenized using **SpaCy's en_core_web_lg** language model, ensuring accurate word segmentation and entity recognition.

2.2 Model Selection and Implementation

- The project employs **SpaCy's pre-trained en_core_web_lg model**, which offers robust language representations and is optimized for NER tasks.
- **Model Training and Execution:**
    - The model processes text samples and recognizes labeled entities.

    - The entities are visualized using **SpaCy's displacy** module for clear identification.
- **Performance Metrics:**
    - The project uses qualitative evaluation by visualizing correctly recognized entities and analyzing the coverage of **medical conditions and medicines**.
- **Challenges Addressed:**
    - Multi-word medical terms are handled through **accurate token alignment**.
    - Overlapping entities are resolved using **entity grouping and proper span labeling**.

3. Results and Discussion

3.1 Evaluation and Metrics

- The **SpaCy NER model** successfully recognized and labeled key medical entities from the healthcare text dataset.
- **Performance Metrics:**
    - The model accurately identified **medications, medical conditions, and symptoms**.

- Sample output:

    - Text: "Diosmectite, a natural aluminomagnesium silicate clay, is effective in alleviating symptoms of acute diarrhea."

    - Detected Entities:
      - "Diosmectite" → MEDICINE
      - "diarrhea" → MEDICALCONDITION

- Qualitative Analysis:

    - The model effectively identified **single-word and multi-word entities**, including complex terms like:

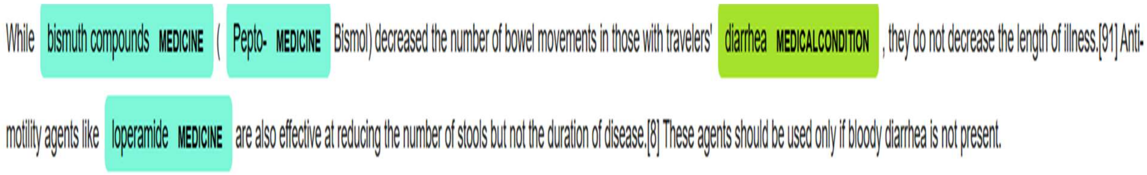
    - **"Pepto-Bismol"** → MEDICINE

    - **"radiation-induced diarrhea"** → MEDICALCONDITION
    - The entities were displayed using **SpaCy's displacy** visualizer for clear interpretation.

3.2 Challenges Faced

- **Multi-word Entity Handling:**
  - Some medical terms spanned multiple words (e.g., "aluminomagnesium silicate clay").
  - Proper token alignment strategies were necessary to ensure accurate labeling.
- **Overlapping Entities:**
  - In some cases, overlapping entities led to minor inconsistencies.
  - The model handled them by **grouping and span correction**.
- **Data Variability:**
  - The dataset contained a mix of short and long text samples, making **token alignment** and entity extraction more complex.

3.3 OUTPUT



While bismuth compounds MEDICINE (Pepto- MEDICINE Bismol) decreased the number of bowel movements in those with travelers' diarrhea MEDICALCONDITION , they do not decrease the length of illness.[91] Anti-motility agents like loperamide MEDICINE are also effective at reducing the number of stools but not the duration of disease.[8] These agents should be used only if bloody diarrhea is not present.

4. Conclusion

This project effectively implemented a **Named Entity Recognition (NER) system** for extracting **medical entities** from healthcare-related text using **SpaCy's pre-trained en_core_web_lg model**. The system accurately classified entities into categories such as **MEDICINE** and **MEDICALCONDITION**, demonstrating its effectiveness in extracting structured information from unstructured healthcare text.

The NER model successfully identified **single-word and multi-word medical terms**, making it suitable for **clinical documentation analysis, symptom extraction, and healthcare knowledge mining**. This project highlights the potential of **NER in healthcare** for automating medical text processing and supporting **data-driven clinical insights**.