Natural Language Processing and Generative AI (CSCI 5800)

Course Project Proposal

Name: Likhith Halkurke Ghanananda Murthy

Student ID: 110971425

Problem Statement

Health technology platforms generate vast amounts of data through patient-doctor interactions, encompassing medical services, prescriptions, and consultations. This data is rich in specialized medical terminology, making it challenging for non-medical professionals to interpret. This project aims to develop a custom Named Entity Recognition (NER) model to identify and extract disease names and their associated treatments from medical texts. It aims to get the accurate disease from the symptoms given by the platform user. Leveraging advanced natural language processing techniques, the NER model will transform unstructured medical notes into structured data. The outcome will enhance the accessibility and utility of medical data on health tech platforms, improving patient care and operational efficiency. The challenge lies in accurately extracting relevant medical entities from unstructured text, which is critical for improving patient care and optimizing healthcare operations.

Background

In the healthcare domain, NER is essential for extracting meaningful information from clinical notes, prescriptions, and other medical documents. The project's significance lies in its potential to enhance data accessibility for healthcare professionals, thereby improving decision-making processes and patient outcomes. The problem is of interest to both academic researchers and industry professionals in healthcare and natural language processing (NLP).

Unique Methodologies that can be used along with NER

- **Adversarial Training for Robustness:** Use of adversarial training methods to improve the robustness of the NER model against noise or ambiguity in medical texts. This could be one of the novel angles since unstructured medical texts might often have spelling mistakes, abbreviations, or inconsistent terminology.
- **Explainable AI:** Incorporate explainability in the model, allowing healthcare professionals to understand why the model made a particular decision. The techniques like attention mechanisms that might help highlight which parts of the text influenced the NER decisions.
- **Contextual Data Augmentation**: Applying medical data augmentation techniques to enhance the training data. For instance, use synonyms to replace medical terms or generate synthetic data using language models.

Data Source

The project will use a dataset consisting of medical records, including clinical notes and prescription data. Kaggle offers a variety of healthcare-related datasets, including medical records, disease information, and treatment data. While some datasets might be limited in size, they can be a good starting point for model development and testing. The National Library of Medicine's (NLM) ClinicalTrials.gov Dataset includes clinical trial summaries that can be useful for

extracting disease and treatment entities. Although it's not as extensive as clinical notes, it provides valuable information in a structured format. The BioNLP Shared Tasks provide datasets specifically for biomedical natural language processing tasks, including named entity recognition, event extraction, and more.

The dataset obtained will be accessed and processed and this includes cleaning, normalization, and annotation to prepare it for model training. This will involve removing irrelevant information, standardizing medical terms, and ensuring consistency across the dataset. A novel approach here can be to use a domain-specific pre-trained model like BioBERT or ClinicalBERT and fine-tune it with the medical dataset that is obtained. Because these models possess pre-existing domain knowledge of medical terms and relationships and this will allow for better performance on medical NER tasks, as they already have some contextual understanding of medical terminology.

Goals and Expectations

The primary goal is to develop an NER model capable of accurately identifying and extracting disease names and their associated treatments from unstructured medical text. Success will be measured using standard NLP metrics such as precision, recall, and F1-score, with a target F1-score of at least 0.85. Additional goals include improving model efficiency and ensuring that the model can handle diverse medical terminology.

The project is expected to yield a robust NER model that can be integrated into healthcare platforms to automate the extraction of critical medical information. This will streamline operations and enhance data-driven decision-making in healthcare settings.

Tools

The project will be implemented using Python, with key libraries including TensorFlow, PyTorch, and spaCy for building and training the NER model. The Jupyter Notebook or a Google Colab environment will be used for development, providing the necessary computational resources. Additionally, healthcare-specific NLP libraries and pre-trained models will be leveraged to fine-tune the NER model. The use of these tools will demonstrate advanced proficiency in NLP techniques and deep learning frameworks, showcasing the knowledge acquired during the MS program.

Products Project will produce

- **Custom NER Model:** A Named Entity Recognition (NER) model specifically tailored to extract disease names and associated treatments from unstructured medical text. This model will leverage advanced NLP techniques to transform unstructured data into a structured format.
- **Data Processing Pipelines:** Methods for cleaning, normalizing, and annotating the medical records dataset to prepare it for model training. This includes removing irrelevant information, standardizing medical terms, and ensuring consistency across the dataset.
- **Comprehensive Project Report:** A final project report documenting the research, implementation process, results, and findings. This report will reflect the project's contribution to the field and demonstrate your mastery of NLP methodologies.