Project Proposal

AIN413
Gülden Olgun

Kemal Şahin
2200765021

March 14, 2024

Contents

1 Introduction **2**

2 Data Description **2**
2.1 Source and Nature of the Data 2
2.2 Features Available 2

3 Methodology **3**
3.1 Data Preprocessing 3
3.2 Model Development 3

Hierarchical Machine Learning Model for Medical Symptom Classification

Kemal Şahin

March 14, 2024

1 Introduction

The objective of this study is to create a hierarchical machine learning model that combines text classification and speech-to-text techniques to accurately analyze and classify medical symptom statements obtained from audio inputs. The ultimate goal of the project is to develop a system that can accurately and contextually comprehend verbal descriptions of medical symptoms.

2 Data Description

This project is based on an essential dataset consisting of 8.5 hours of carefully selected audio recordings along with textual annotations. These recordings capture a wide variety of medical symptoms expressed by different people. The main components of the dataset are audio files and the textual metadata that goes with them.

2.1 Source and Nature of the Data

The dataset used in this research comes from Figure Eight, who is now Appen and is well-known for having powerful datasets in a variety of fields. This dataset can be accessed on Kaggle.

Thousands of audio clips make the entire dataset, each carefully matched with a textual transcription that describes the audible symptom, so representing a wide range of possible medical conditions.

Acknowledgements: This dataset was developed and made available by Figure Eight. It can be accessed for public use, along with guidelines on replicating similar datasets, at https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent/.

2.2 Features Available

Each dataset entry comprises two principal components:

- **Audio File:** The raw audio recording of a spoken symptom description, serving as the primary input for the speech-to-text component of the model.
- **Textual Annotation:** The corresponding textual transcription of the audio content, intended for validating the speech-to-text output and serving as the basis for text classification.

Furthermore, each recording has metadata included to support a thorough analysis and preprocessing stage, such as the recording quality, perceived clarity, and other important annotations.

3 Methodology

This section outlines the thorough approach used in this study to build a hierarchical machine learning model that can recognize and categorize audio recordings of medical symptoms. The procedure is divided into three distinct stages: model development, evaluation, and data preprocessing. Each stage is essential to the project's overall goal.

3.1 Data Preprocessing

The quality and consistency of the dataset are crucially important, and they have a direct impact on the model's performance during the data preprocessing stage. The actions consist of:

1. **Clustering for Data Cleaning:** Utilizing clustering algorithms to identify and seperate poor-quality audio files and mislabeled data, which are then reviewed and corrected or removed from the dataset.

2. **Feature Encoding:** Converting categorical metadata into a format that can be processed by machine learning algorithms; in this case, one-hot encoding of nominal features and proper scaling of all numerical features will be used.

3. **Text Embedding:** Applying advanced text embedding techniques, such as BERT embeddings or alternative encoders, to transform the textual transcriptions obtained from the speech-to-text model into a numerical format suitable for the following text classification. This step is crucial for capturing the semantic richness of the symptom descriptions, enabling more precise and advanced classification.

3.2 Model Development

The hierarchical machine learning model is designed and implemented during the model development phase. It is organized as follows:

1. **Speech-to-Text Conversion:** The initial layer of the model employs a speech-to-text engine to transcribe the audio recordings into textual data.

2. **Text Classification:** The subsequent layer utilizes a text classification model to categorize the transcribed texts into predefined medical symptom categories.

3. **Integration and Tuning:** Both layers are integrated into a compatible pipeline, with careful tuning of parameters and optimization based on performance metrics.