## SymptoSense : NLP Based Symptom Classifier

***Synopsis submitted to***

***Shri Ramdeobaba College of Engineering & Management, Nagpur in partial fulfillment of requirement for the award of***

***degree of***

## Bachelor of Engineering

*In*

## COMPUTER SCIENCE AND ENGINEERING

## (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

*By*

### Mr. Shreyas Raut

### Mr. Samiran Gupta

### Mr. Shubham Shivhare

### Mr. Vedant Waghale

*Guide*

### Dr. Amit Pimpalkar



**Computer Science and Engineering (AI ML)**

### Shri Ramdeobaba College of Engineering & Management, Nagpur

**440013**

**SymptoSense : NLP Based Symptom Classifier**

**Project Definition:**

Developing an accurate system for the classification of medical symptoms from text inputs in the field of Natural Language Processing (NLP). The objective is to create a model that can effectively analyze and categorize patient-reported symptoms, in written form , to aid healthcare professionals in diagnosis and treatment planning.

**Project objectives:**

The primary objective of this project is to develop a conversational agent for addressing common medical symptoms using a dataset comprising over 6000 transcripts. Each transcript, created by individual human contributors based on a given symptom, is intended to train the conversational agent in the medical field .The specific objectives encompass exploratory data analysis, including a detailed exploration of the dataset structure, label correction, identification and handling of poor-quality transcripts and data preprocessing, including text tokenization and a proper train-validation-test split, the dataset is prepared for training a NLP model.

The goal is to build a model capable of understanding and responding to user queries related to medical symptoms, deploying it in a medical environment.The project aims to enhance the interaction between users and healthcare resources by providing an effective and user-friendly conversational agent for medical assistance.

Following the training of the model, a thorough evaluation is conducted, employing performance metrics like accuracy and precision. The iterative nature of the project involves fine-tuning of the model.

**Proposed Plan of Work**:

**Project initiation:**

* Create a robust symptom classification system that can process textual descriptions.
* By leveraging NLP techniques, we can assist healthcare professionals in identifying symptoms accurately and efficiently.

**Scope:**

* **Text-Based Symptoms:** Analyzing symptom descriptions provided in text format (e.g., patient notes, medical reports).

**Data Collection and Preprocessing :**

* **Data collection:** Gather relevant datasets containing symptom-related text . Explore medical literature, patient records, and publicly available healthcare databases and our source , <https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent> .
* **Data Preprocessing:** Clean and normalize the data , Tokenize text data.

**Feature Extraction :**

* **Text Features:**
  + TF-IDF (Term Frequency-Inverse Document Frequency): To represent text data.
  + Word Embeddings (Word2Vec, GloVe , FastText): To capture semantic relationships between words.
  + BERT (Bidirectional Encoder Representations from Transformers): For contextualized embeddings.

**Model Selection and Training (Algorithm) :**

* **Text Classification Models :**
  + - Naive Bayes, SVM, or deep learning models (LSTM, BERT,Transformers).
* Split data into training, validation, and test sets.
* Train and fine-tune models using selected features.

**Evaluation Metrics :**

* **Define evaluation metrics :**
  + Accuracy, precision, recall, F1-score, AUC-ROC.
  + Consider the medical context and prioritize minimizing false negatives (missing symptoms).

**Model Evaluation and Validation :**

* Evaluate models on the validation set.
* Fine-tune hyperparameters.
* Validate models on a separate test set to assess generalization.

**User interface design :**

* Select a frontend framework or library (e.g., react) and set up the development environment.
* Design the user interface (UI) for data input (user interaction) and results display.
* Choose a backend framework (e.g., flask, django) and set up the development environment.

**Methodology:**

**Data Understanding and Exploration:**

* Load and Inspect the Dataset :

o Load the recordings-overview.csv file into a Pandas DataFrame.

o Understand the structure of the dataset, including columns, data types, and basic statistics.

* Explore Transcripts:

o Examine a sample of transcripts to understand the language used and identify any issues with quality, grammar, or completeness.

* Analyze Label Distribution

o Check the distribution of medical symptoms in the dataset to ensure it's representative and balanced

**Data Cleaning:**

* Transcript Quality Check:

o Implement checks for poor-quality transcripts. Consider removing or flagging transcripts with incomplete sentences, unclear descriptions, or irrelevant content.

* Duplicate Removal:

o Check for and remove any duplicate entries to maintain data integrity.

* Data Balancing:

o Address any class imbalances by oversampling or undersampling, ensuring each symptom has sufficient representation.

**Data Preprocessing:**

* Text Tokenization:

o Tokenize the transcriptions into a format suitable for NLP.

* Train-Validation-Test Split:

o Split the dataset into training, validation, and test sets for model training and evaluation.

**Model Development:**

* Choose Model Architecture:

o Select a NLP model architecture, such as LSTM or BERT.

* Model Training

o Train the model using the preprocessed data. Monitor training performance and adjust hyperparameters if needed.

**Model Evaluation:**

* Performance Metrics:

o Define evaluation metrics such as accuracy, precision, recall, and F1 score.

* Test Set Evaluation:

o Evaluate the model on the test set to assess its overall performance.

**Fine-Tuning and Iteration:**

* Model Fine-Tuning:

o Based on the evaluation results, fine-tune the model to improve performance.

* Iterative Improvement:

o Iterate through the cleaning, preprocessing, and modeling steps based on feedback and additional insights.

**Technology:**

**Tokenization and Sentence Segmentation:**

* NLTK's word\_tokenize and sent\_tokenize for tokenization.

**Text Normalization:**

* NLTK's WordNetLemmatizer for word root form.

**Text Cleaning:**

* String and re for basic string manipulation and regex.

**Feature Extraction:**

* TfidfVectorizer for TF-IDF features,word embedding ( Word2Vec, GloVe , FastText).

**DL Models for Text Classification:**

* Choose deep learning models (e.g., LSTM, BERT, Transformers).

**Handling Imbalanced Classes:**

* SMOTE for addressing class imbalance.

**Visualization:**

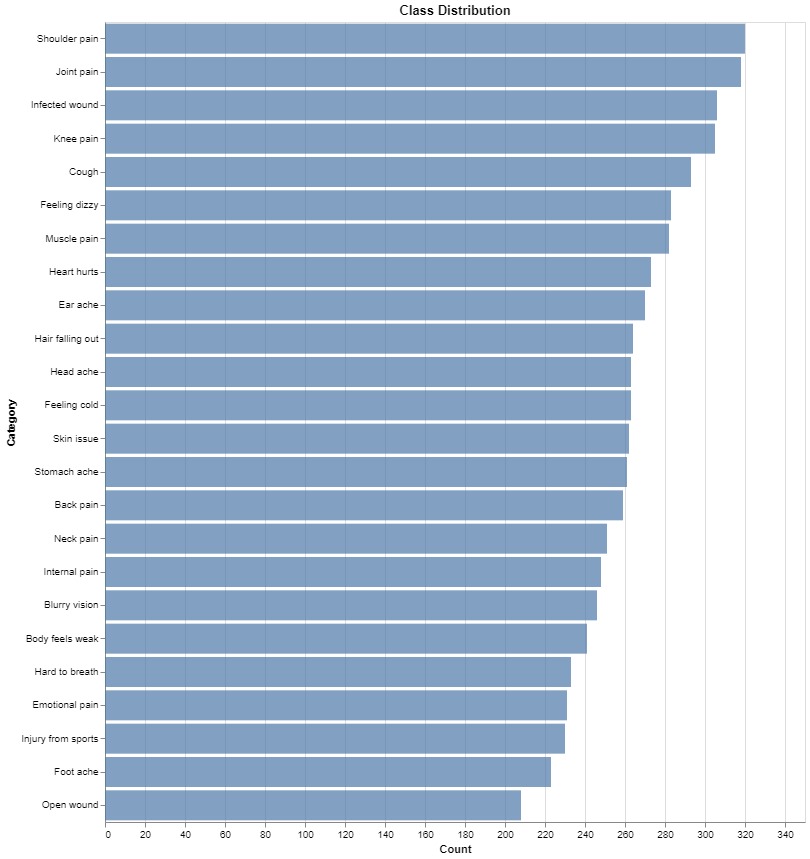
* Matplotlib and seaborn for result visualization.

**Data and Hardware:**

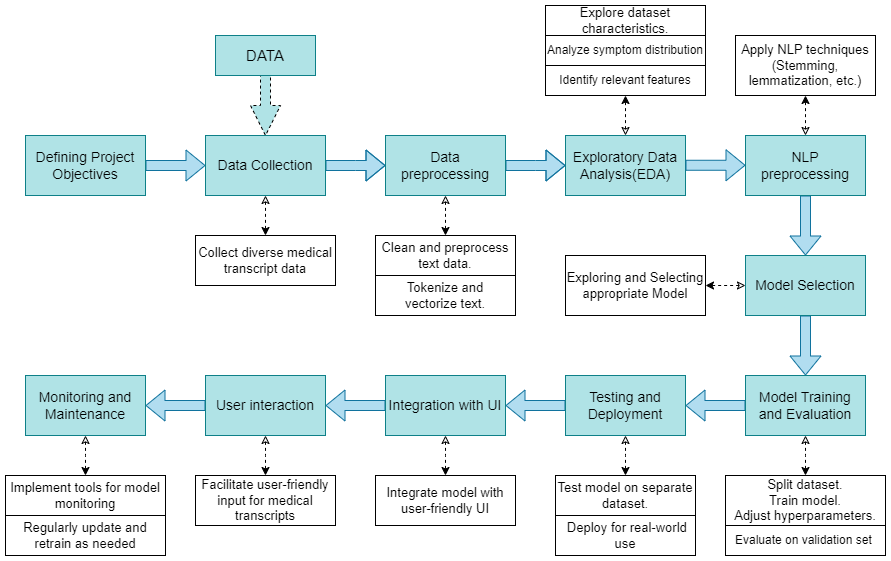
* Kaggle for dataset; GPU for enhanced performance.

**Dataset Overview:**

**Dataset Description :** The dataset comprises a CSV file containing 6000 transcripts related to healthcare interactions. Each transcript represents a patient's textual input, capturing various aspects of their healthcare experiences, concerns, and interactions with healthcare professionals. Additionally, the dataset includes prompts derived from the patient's input. These prompts aim to encapsulate key themes, inquiries, or issues raised by the patients during their interactions. The dataset serves as a valuable resource for healthcare professionals, researchers, and analysts interested in understanding patient perspectives, healthcare communication patterns, and the overall dynamics for healthcare encounters**.**



**WorkFlow:**



**Project Scope:**

The project scope encompasses the development of a conversational agent for addressing a diverse range of common medical symptoms.The conversational agent is designed to understand and respond to user queries related to medical symptoms, enhancing user interaction with healthcare information.

**Details of the Group Members:**

| **Roll No.** | **Name of Students** | **Name of the Guide** |
| --- | --- | --- |
| E - 57 | Mr. Shreyas Raut | Dr. Amit Pimpalkar |
| E - 54 | Mr. Samiran Gupta |  |
| E - 58 | Mr. Shubham Shivhare |  |
| E - 60 | Mr. Vedant Waghale |  |

Approved By :-

Dr. Amit Pimpalkar Prof. Priya Parkhi

(Guide) (Project Coordinator)