**Technical Design Document (TDD)**

**Title: Feature Extraction from Medical Journals**

|  |  |
| --- | --- |
| **Mentors*:*** | **Authors:** |
| *Pavani*  *Raju*  **Mentor (Client):**  *Sharat Chandra*  *Sandeep* | *Mohammed Zaid Habibulla* |

**Table of Contents**

[1. Introduction 4](#_Toc89779544)

[1.1 Purpose 4](#_Toc89779545)

[1.2 Scope 4](#_Toc89779546)

[1.3 Document Organization 4](#_Toc89779547)

[1.4 Audience 4](#_Toc89779548)

[2. System Overview 5](#_Toc89779549)

[2.1 Context 5](#_Toc89779550)

[2.2 Product Feature 5](#_Toc89779551)

[2.3 Technologies Used 5](#_Toc89779552)

[3. Model Architecture 5](#_Toc89779553)

[3.1 Model Training 5](#_Toc89779554)

[3.2 Transferable Model and Data Preparation 5](#_Toc89779555)

[3.3 Run the model on our data 5](#_Toc89779556)

[3.4 Model Evaluation 6](#_Toc89779557)

[4. Approach 6](#_Toc89779558)

[4.1 Extract the data 6](#_Toc89779559)

[4.2 Data Preparation and EDA 6](#_Toc89779560)

[4.3 Transferable model 6](#_Toc89779561)

[5. Assumption and Constraints 6](#_Toc89779562)

[6.1 Project Assumptions 6](#_Toc89779563)

[6.2 Project Constraints 6](#_Toc89779564)

[5.3 References 6](#_Toc89779565)

**Revision History**

|  |  |  |
| --- | --- | --- |
| Version | Date | Description |
| 1.0 | 02/12/2021 | Initial Draft |
|  |  |  |

# Introduction

## 1.1 Purpose

 Manual annotation and feature generation by biomedical experts are inefficient because they involve a complex process and require expensive and time-consuming labour. Therefore, efficient and accurate natural language processing (NLP) techniques are becoming increasingly important for use in computational data analysis, and advanced text mining techniques are necessary to automatically analyse the biomedical literature and extract useful information from texts. To bridge the gap between academic development and industrial requirements, we are developing a web application that recognizes and extracts the entities from the content which helps to minimize time commitments from domain experts and the manual efforts on researching content.

## Scope

* Our major goal is recognizing Diseases, biomedical or Chemical composition names of drugs by  named entity recognition (NER) which is an important task used to extract information and to recognize the Medical entities from Medical Journals.
* To maximise the utilisation of free-text medical journals, we focused on a particular sub-task of disease names,drug and it’s compositions information extraction and developing a dedicated named-entity recognition web application.

## 1.3 Document Organization

This document is organized into the following section:

|  |  |
| --- | --- |
| Introduction | Provides information related to the document |
| System Overview | Describes the approach, architectural goals and constraints, Guiding principles |
| Model Architecture | Describes model architecture |
| Assumptions and Constraints | Details various assumptions made during design and development of the Online Screening tool |

## Audience

The intended audiences for this document are: -

* Innodatatics Inc.
* The project development team
* Mentors

# System Overview

## 2.1 Context

Manual annotation and feature generation by biomedical experts are inefficient because they involve a complex process and require expensive and time-consuming labour. Therefore, efficient and accurate natural language processing (NLP) techniques are becoming increasingly important for use in computational data analysis, and advanced text mining techniques are necessary to automatically analyse the biomedical literature and extract useful information from texts. To bridge the gap between academic development and industrial requirements, we are developing a web application that recognizes and extracts the entities from the content which helps to minimize time commitments from domain experts and the manual efforts on researching content.

## 2.2 Product Feature

The major feature of the ***Feature Extraction from Medical Journals*** will be the following: -

* + Dosage – The total amount of the drug administered
  + Drug – Generic or Brand name of the medication
  + Duration – The length of time that the drug was prescribed for
  + Form - A particular configuration of the drug which it is marketed for use
  + Frequency – The dosage regimen at which the medication should be administered
  + Route – The path by which the drug is taken into the body
  + Strength – The amount of drug in a given dosage

## Technologies Used

The model was implemented in python – jupyter Notebook. Libraries used are med7 for extraction of features, PDFminer for the text extraction. Anaconda for the environment setup.

# Model Architecture

## 3.1 Model Training

**Med7 –** The model was initially trained on MIMIC-III dataset for extracting medical entities from clinical text and further the model was also tested and trained on UK-Cris and NHS datasets

## 3.2 Transferable Model and Data Preparation

Med7 is a transferable clinical NLP model for electronic health records, compatible with spaCy, for clinical named entity recognition tasks. We have used this model to extract features from our medical journals.

## 3.3 Run the model on our data

Med7 model has been used to extract features like drug, duration, dosage, form, frequency, route and strength from the medical journals

## 3.4 Model Evaluation

To evaluate the accuracy of the features extracted by the model on our data, we need to annotate the documents with the help of a subject domain expert

# Approach

## 4.1 Extract the data

Data, from the medical journals which are in pdf format, is extracted using PDFminer

## 4.2 Data Preparation and EDA

Using ***NLTK* ,** stop words are removed, lemmatization is done. Word cloud is prepared to do EDA

## Transferable model

Using ***Med7*** model, required features are extracted

# Assumption and Constraints

## 6.1 Project Assumptions

* Data is in an extractable format like pdf
* Data consists of features being extracted

## 6.2 Project Constraints

* Maximize the quantity of features extracted from the medical journals
* Minimize incorrect feature detection
* Minimize the run time of the model training

## References

<https://pdfminersix.readthedocs.io/en/latest/>

<https://pubmed.ncbi.nlm.nih.gov/34412834/>

<https://github.com/kormilitzin/med7>