Anusha Kanagala Z1901136
Snehal Utage Z1888637

# CSCI 642 PROJECT PROPOSAL

**Name:** Anusha Kanagala (Z1901136), Snehal Utage (Z1888637)

## Problem Significance:

Clinical trials are the medical research with volunteers [5]. They are the core component for discovering new drugs, vaccines, devices, procedures, and treatments for diseases. It helps in studying the efficacy and safety of a new drug and treatment in people. They play a significant role in translating scientific research into the practice of medical outcomes. During the Covid-19 pandemic, we have seen many clinical trials conducted for the development of vaccines. Clinical trials data is available publicly and monitored by many clinical registries and institutions. But many clinical trials are being conducted worldwide, and it is difficult for researchers and participants to search for the appropriate and relevant clinical trial to participate for a specific health condition or category. Data extraction or classification from original study reports or these big platforms (large datasets) is time-consuming and error prone. Clinical trials data have abstract clinical narratives that give information about the patient's history and treatments. Hence, we are using text classification that will help identify the health conditions studied in a particular clinical trial from its description text. We are using the data from Dimension.ai, which provides detailed information about many clinical trials for a specific period.

Text classification brings automation and simplification to the search and retrieves information from datasets, and it standardizes the platform. It improves user experience by simplifying navigation and retrieving the relevant information. Text Classification is a task of automatically classifying/labeling the documents into one or more predefined categories. It is one of the fundamental tasks of the Machine Learning (ML) algorithm for Natural Language Processing (NLP). It is a critical component in many NLP applications, such as sentiment analysis, relationship extraction, and spam detection. This classification will help the volunteers to search for the appropriate clinical trial to participate in, if a researcher wants to identify any new studies conducted in a clinical trial, then they can get the details about it. This predictive analysis will help the researchers and the participants of clinical trials to search the relevant clinical trials quickly. It will help researchers to answer different questions related to health conditions. It will help to retrieve the relevant clinical trials details, what procedures were followed.

## Research hypotheses:

As part of this project, we will predict the health categories/conditions, and cancer types from the description of clinical trials. We will build a Machine Learning model that will automatically classify and predict the health category/condition studied in the clinical trial using Text Classification.

**Related work:**

Several studies have analyzed clinical trials data using supervised machine learning algorithms. Work is done in screening and classifying the eligibility criteria, clustering the eligibility criteria, classification of cancer, diabetes clinical trial data using Deep Learning, multilabel classification of clinical data, and many more. Study of clinical trials text and its classification is developed through several approaches such as eligibility screening approach [9,8,14,18]; machine learning approach [6,7,15]; deep neural network [6,8,15], convolutional neural network [8,15] and approaches to fine grained document clustering [6] and many others.

Jasmir et. al used the supervised ML such as K-Nearest Neighbor(k-NN), Decision Tree, SVM, Random Forest, Deep Neural Network, and Fine-Grained Document Clustering to create classification model for Cancer clinical trial documents collections that improved the computational value. In their study they found Fine-Grained clustering improved the computation value of classification [6,17].

In other study K. Zhang et. al automated the classification of eligibility criteria in clinical trials to facilitate the patient-trial matching for a specific population of patients. They used regex model, Bag-of-words ML classifier, and ML classifier with named entity recognition (NER). As per their results the addition of NER didn't improve the classification results [7].

C. Chuan proposes an active deep learning approach to automatically classify clinical trial eligibility criteria. The experimental results showed that active CNN performed significantly better than the (k-NN), method and word2vec successfully learns meaningful embeddings in criteria [8].

Y. Ni et al discuss the development of an automatic screening eligibility algorithm to identify patients who meet the core eligibility characteristics of oncology clinical trials [9]. An automated approach to clustering clinical trials with similar eligibility features was studied by T. Hao et. al [10]. K. Zheng et. el proposed a classifier for short text eligibility criteria based on ensemble learning [14].

Xu Rong et. al developed a novel automated approach to structure random clinical trials (RCT) abstracts by combining text classification and Hidden Markov Modeling (HMM) techniques. The text classification methods were able to effectively categorize sentences in RCT abstracts when combined with the HMM technique [11]. ML techniques such as SVM and linear classifiers was used to automate labeling sentences in an abstract with subheadings [12].

Richard E. reviews how the natural language processing techniques of Term-Frequency Inverse-Document-Frequency (TF-IDF) combined with the supervised machine learning model of Support Vector Machines (SVM) and word embedding approaches such as word2vec can be used to

categorize/label protocol deviations across multiple therapeutic areas. NLP is a key tool that will lead to more data driven decisions in clinical trial operations [13].

Atal I. et. el. developed a knowledge-based classifier that allow automatic mapping of the health conditions (multi labels) studied in registered clinical trials to categories of diseases and injuries from the Global Burden of Diseases (GBD) 2010 study, it relies on the UMLS® knowledge source (Unified Medical Language System®) and heuristic algorithms for parsing data [16].

## Data:

For training and testing our ML classification model, we are using the clinical trial data available from Dimensions.ai [1][2]. Dimensions is a project that maps the entire research lifecycle and provides detailed research information. It is a comprehensive database of collection of linked data in a single platform, from grants, publications, datasets, and clinical trials to patents and policy documents. Dimensions currently have data-related Clinical trials 641k.

The Clinical trials dataset which we are using has 32,728 records and 57 different features. Few features include Trial ID, Title, Abstract, Gender, Phase, Condition, Sponsor, Funder Country, and Categories. Categories include Fields of Research (FoR), RCDC (Research, Condition, and Disease Categorization), HRCS (Health Research Classification System) – HC (Health Categories) and RAC (Research Activity Classification), ICRP (International Cancer Research Partnership) Cancer Types, ICRP CSO (Common Scientific Outline) [3][4]. We are focusing on the title text and HCRS_HC, ICRP Cancer Types categories.

## Methods:

The Clinical Trial dataset has about 57 features, the title, abstract, categories, conditions, cancer types, etc. We are focusing on text from the title and abstract. For the target, we will consider the Condition, HRCS HC, and ICRP Cancer type categories. The input to our ML classifier model is the clinical text, and the target feature is the either health categories, conditions, or cancer type.

We will train different supervised machine learning classifier models – Naïve Bayes', Logistic Regression, SVM, K-Nearest Neighbors. For each model, we will split the dataset into two groups the Training Dataset and the Testing Dataset. We will use 80% for training and 20% data for testing. We will find the different model's accuracy, precision, recall, and F1-score and compare them.

We will follow below steps:

- In the first step, we will import the libraries and load the dataset.
- Next, we will do Exploratory data analysis on the data for understanding the distribution of each health condition, category, and cancer type. We will check for any missing values and drop those records. Visualize the class and word distributions.
- The next step is text preprocessing - stop word removal, stemming, lemmatization using NLTK library functions.
- Supervised learning algorithms will require a category label for classification in the training set. In this case, the category label we will focus on is the Conditions, HRCS HC, and ICRP Cancer Types.
- Later we will convert the text to tokens using Tokenization (TF-IDF) with Scikit-Learn.
- Then we will train different classifier models available in Scikit-Learn on training data and perform predictions on the test data.
- We will compare the results obtained from different classifier models.

We will use the Google Colab server and python programming language for the development of code. We will use the Scikit-Learn, NLTK, NumPy, Pandas, Matplotlib libraries in our project.

**Innovation:**

We are creating an ML model that will automatically predict the health category studied in the clinical trial based on the description of the clinical trial. We are exploring the Dimensions.ai clinical trial dataset to classify the text and help in improving the search. We are comparing the results of different ML classifier algorithms. It will help the participants and researchers of clinical trials to search the trials related to health conditions and categories. It might also help the Insurance authorities to verify the trial if it is related to a health category. It might help them in making decisions about claims.

**Evaluation:**

For evaluating the results obtained from different classifier models, we will consider the Accuracy, Precision, Recall, and F1-score metrics. Based on these metrics, we can identify which classifier model predicts the health conditions/categories/cancer types accurately.

Anusha Kanagala Z1901136
Snehal Utage Z1888637

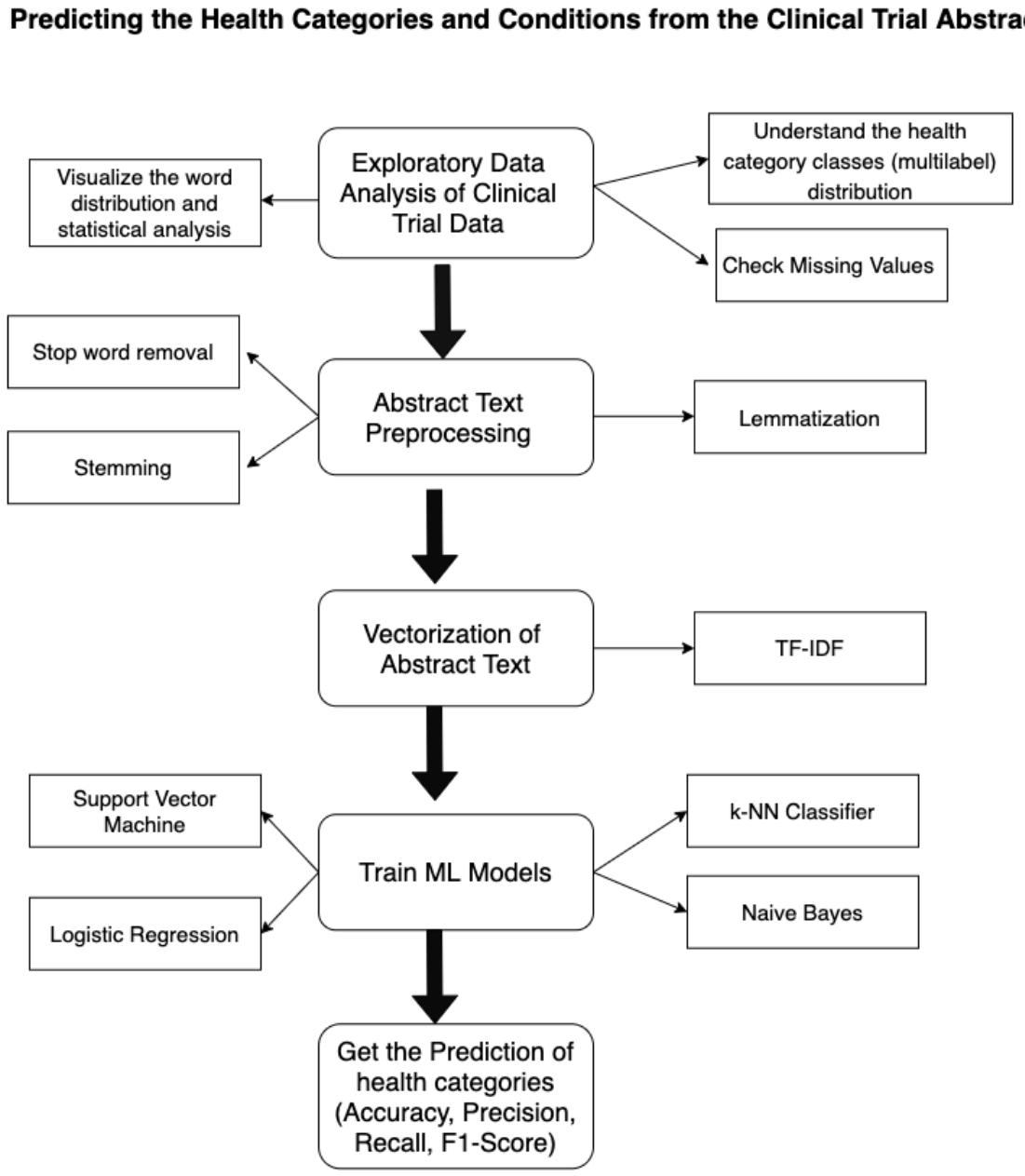
**Time plan:**

| Week/ Task | Task Progress | Task Progress | Task Progress | Task Progress | Team Member |
|---|---|---|---|---|---|
| **Week 1** | Initial Exploratory Data Analysis, visualization of class labels. | | | | **Snehal Anusha** |
| **Week 2** | | Pre-processing of dataset. Vectorization of input text to be passed to ML model. | | | **Snehal** |
| **Week 3** | | | Train supervised ML model for Text classification based on Health Category, Conditions, and Cancer Type. | | **Anusha Snehal** |
| **Week 4** | | | | Evaluation of Accuracy of the models and compare the results of models. | **Anusha** |

**Expected results:**

We will compare the results of the classification algorithms k-NN, Support Vector Machine, Naïve Bayes', and Logistic regression.

The overall steps we will follow are (developed using diagrams.net [20])-



**References:**

[1] Why did we build Dimensions?
https://www.dimensions.ai/why-dimensions/

[2] Dataset uploaded on Dropbox -
https://www.dropbox.com/s/8jjzdcbtgyx9lht/Dimensions-Clinical-Trial-2019-11-01_00-39-34.xlsx?dl=0

[3] Which research categories and classification schemes are available in Dimensions?
https://plus.dimensions.ai/support/solutions/articles/23000018820-which-research-categories-and-classification-schemes-are-available-in-dimensions-

[4] Clinical Trials, Clinical Trial Fields
https://docs.dimensions.ai/dsl/datasource-clinical_trials.html

[5] National Institute for Health
https://www.nih.gov/

[6] J. Jasmir, S. Nurmaini, R. F. Malik, and D. Zaenal, "Text Classification of Cancer Clinical Trials Documents Using Deep Neural Network and Fine-Grained Document Clustering," Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019), 2019

[7] K. Zhang and D. Demner-fushman, "Automated classification of eligibility criteria in clinical trials to facilitate patient trial matching for specific patient populations," vol. 0, no. June 2016, pp. 1–7, 2017

[8] C. Chuan, "Classifying Eligibility Criteria in Clinical Trials Using Active Deep Learning," 2018 17th IEEE Int. Conf. Mach. Learn. Appl., pp. 305–310, 2018.

[9] Y. Ni et al., "Increasing the efficiency of trial-patient matching: Automated clinical trial eligibility Pre-screening for pediatric oncology patients. Clinical decision-making, knowledge support systems, and theory," BMC Med. Inform. Decis. Mak., vol. 15, no. 1, pp. 1-10, 2015

[10] T. Hao, A. Rusanov, M. R. Boland, and C. Weng, "Clustering clinical trials with similar eligibility criteria features," J. Biomed. Inform., vol. 52, pp. 112–120, 2014

[11] Xu, Rong et al. "Combining text classification and Hidden Markov Modeling techniques for categorizing sentences in randomized clinical trial abstracts." AMIA ... Annual Symposium proceedings. AMIA Symposium vol. 2006 (2006): 824-8.

[12] McKnight L, Srinivasan P. Categorization of sentence types in medical abstracts. Proceedings of the 2003 AMIA conference.2003

[13] Richard, E., Reddy, B. Text Classification for Clinical Trial Operations: Evaluation and Comparison of Natural Language Processing Techniques. Ther Innov Regul Sci 55, 447–453 (2021). https://doi.org/10.1007/s43441-020-00236-x

[14] Zeng K, Pan Z, Xu Y, Qu Y "An Ensemble Learning Strategy for Eligibility Criteria Text Classification for Clinical Trial Recruitment: Algorithm Development and Validation"
JMIR Med Inform 2020;8(7): e17832 URL: https://medinform.jmir.org/2020/7/e17832
DOI: 10.2196/17832

[15] Wang, Y., Sohn, S., Liu, S. et al. A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inform Decis Mak 19, 1 (2019). https://doi.org/10.1186/s12911-018-0723-6

[16] Atal, I., Zeitoun, JD., Névéol, A. et al. Automatic classification of registered clinical trials towards the Global Burden of Diseases taxonomy of diseases and injuries. BMC Bioinformatics 17, 392 (2016). https://doi.org/10.1186/s12859-016-1247-7

[17] Jasmir, Jasmir, Siti Nurmaini, and Bambang Tutuko. 2021. "Fine-Grained Algorithm for Improving KNN Computational Performance on Clinical Trials Text Classification" *Big Data and Cognitive Computing* 5, no. 4: 60. https://doi.org/10.3390/bdcc5040060

[18] Tseo Y, Salkola M I, Mohamed A, et al. Information extraction of clinical trial eligibility criteria 2020; arXiv preprint arXiv:2006.07296.

[19] Scikit-Learn for Text classification
https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

[20] NLP Tutorial for Text Classification in Python
https://medium.com/analytics-vidhya/nlp-tutorial-for-text-classification-in-python-8f19cd17b49e

[21] Draw flow diagrams
https://app.diagrams.net/