Clinical Trial Matching and Medical Transcriptions using BERT and XLNet

Sandhya Arumugam Karunanithy
arumugam@purdue.edu
Purdue University
Indiana, USA

Steve Cho
cho512@purdue.edu
Purdue University
Indiana, USA

ACM Reference Format:
Sandhya Arumugam Karunanithy and Steve Cho. 2023. Clinical Trial Matching and Medical Transcriptions using BERT and XLNet. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Clinical trials and medical systems play a critical role in advancing the medical field and improving patient care. One key challenge in this domain is efficiently matching patients with appropriate clinical trials based on their individual profiles and trial requirements. Existing methods for identifying suitable patients are often manual and time-consuming, which limit their effectiveness and scalability.

Our project aims to address this problem to develop a Natural language model that can effectively match patients to suitable clinical trials[1] based on patient profiles and Clinical trial descriptions.

Our project tackles two different tasks:

- 1. To fetch top 'k' patient profiles which are suitable for given clinical trial description[5] using BERT (Bidirectional Encoder Representations from Transformers)model
- 2. To identify medical symptoms based on the patient's descriptive text using XLNet[2].

The potential benefits of our approach include reduced manual effort, increased accuracy in matching patients with suitable trials, and ultimately, improved patient outcomes.Our work builds on existing research in natural language processing, clinical trial matching, and medical text analysis. By incorporating state-of-the-art language models like BERT and XLNet, we aim to advance the field and create a more efficient patient-trial matching system.

2 EXPLORATORY DATA ANALYSIS - CLINICAL TRIALS

The below graphs Figure1, Figure2, Figure3 show the distribution of gender, age, and top 15 medical disorders observed in the patients.

The graphs Figure4 and Figure5 show the distribution of study types and enrollment sizes in clinical trial data.

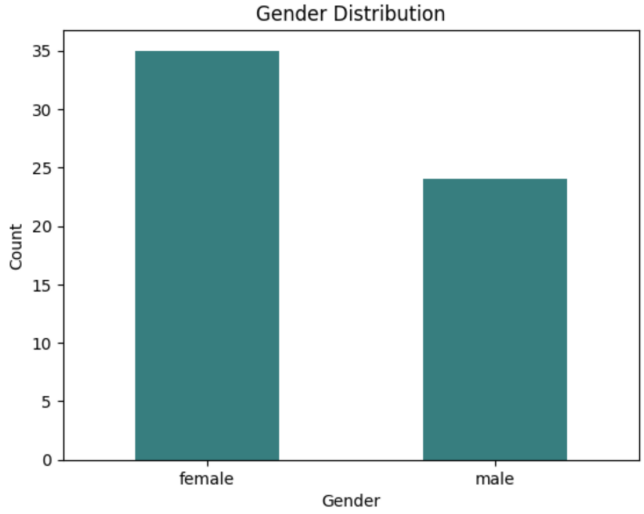

Figure 1: Gender distribution

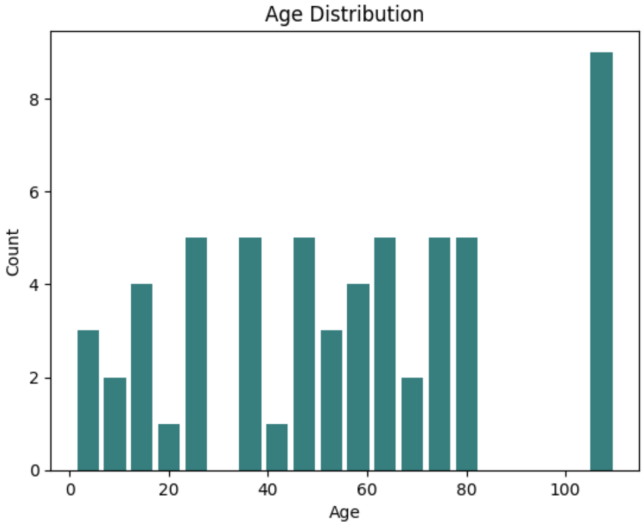

Figure 2: Age distribution

3 EXPLORATORY DATA ANALYSIS - MEDICAL SYMPTOMS

The below table shows the distribution of data for each corresponding medical symptom. As seen in the table, the number of classes is fairly distributed, indicating that the data set is well-balanced.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

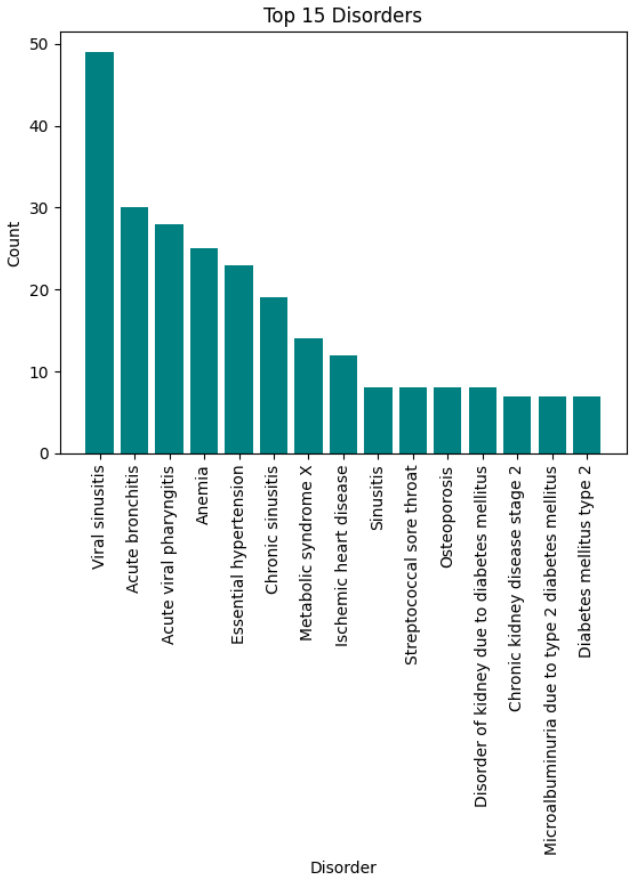

Figure 3: Top 15 Medical Conditions Observed

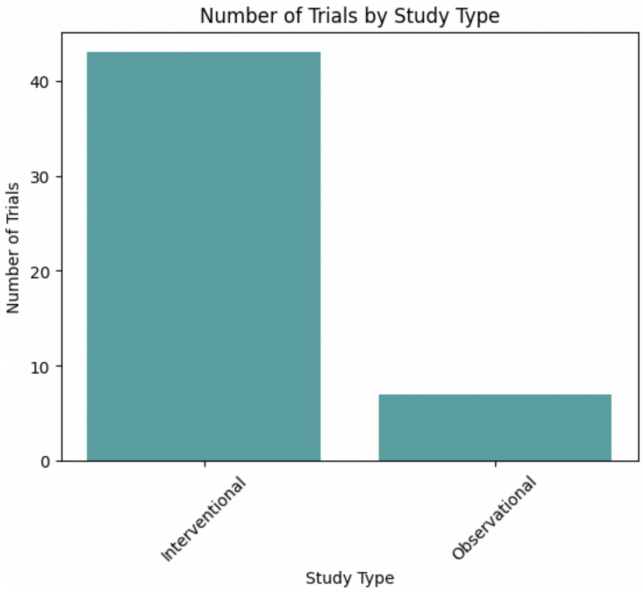

Figure 4: Distribution of type of Study

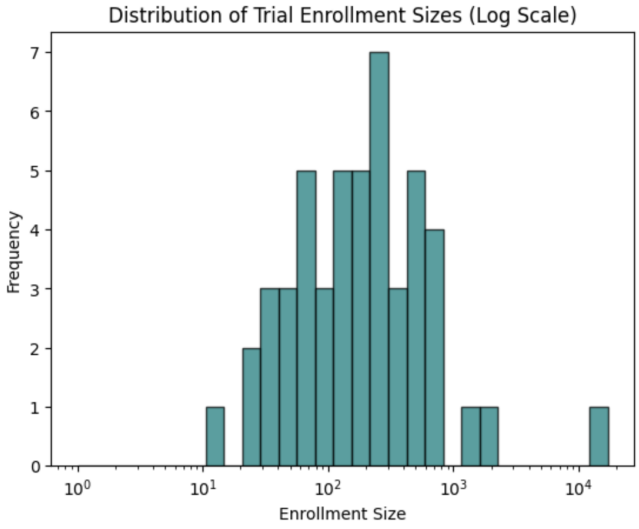

Figure 5: Enrollment size frequency for all trials

Distribution of Medical Symptoms	
Symptom Name	Value Count
Feeling cold	263
Skin issue	262
Stomach ache	261
Back pain	259
Neck pain	251
Internal pain	248
Blurry vision	246
Body feels weak	241
Hard to breath	233
Emotional pain	231
Injury from sports	230
Foot ache	223
Open wound	208

Distribution of Medical Symptoms Continued	
Symptom Name	Value Count
Acne	328
Shoulder pain	320
Joint pain	318
Infected wound	306
Knee pain	305
Cough	293
Feeling dizzy	283
Muscle pain	282
Heart hurts	273
Ear ache	270
Hair falling out	264
Head ache	263

4 METHODS - BERT

The Technical Approach for BERT is as follows:

- Data extraction: We first fetch the clinical trial data and the patient profile data as per the process mentioned in the Dataset section.
- Data pre-processing: We then tokenized the text and removed stop words. We focused on processing eligibility criteria (extracting age and gender) and descriptions of the clinical trials and the conditions of the patients and converted them as input features suitable for BERT.
- BERT Model: We then loaded the pre-trained BERT Model and computed BERT embeddings for the pre-processed conditions of the patients and descriptions of the clinical trials.We chose this approach because BERT has shown strong performance in a variety of natural language processing tasks, including semantic similarity, which is essential for patient-trial matching.
- Calculating Similarities for the match: We then calculated Cosine similarities between the patient conditions and trial descriptions with BERT embeddings.
- Top recommendations: We filtered clinical trials based on the patient's gender and age and sorted the filtered trials based on the calculated cosine similarities and returns the top k recommended trials for each patient. Currently, we are calculating the top 3 recommendations for each patient.

Our approach addresses the problem of efficiently matching patients with clinical trials by using BERT's ability to understand semantic relationships. However, there are challenges associated with using BERT, such as its computational complexity and the need for sufficient data to generate accurate embeddings. We opted to use a pre-trained BERT model to overcome data sparsity and reduce overall training time.

5 METHODS - XLNET

In addition to the BERT model, we also employed the XLNet model for the task of identifying medical symptoms from patients' descriptive text.

The Technical Approach for XLNet is as follows:

- Data pre-processing: We split the data set into train, test, and validation sets. For tokenization and input formatting, we use a pre-trained XLNet tokenizer imported from transformers. The tokenizer requires SentencePiece, which implements subword units and a unigram language model with direct training from given input sentences. The maximum length of tokens was measured by obtaining the length of input ids of each input sentence using the tokenizer's encoding method[3].
- Constructing data loader: After obtaining tokenized version of each training, testing, and validation data set, we construct the data loader of each data set. For each data set, we converted input ids, attention masks, and encoded labels into lists of tensors.
- XLNet Model: We use a pre-trained XLNet model for sequence classification provided by HuggingFace with given 25 labels. We also incorporate a linear scheduler for AdamW optimizer for learning rate scheduling.
- Training and testing: We finally train the model with the given training data loader to obtain accuracy, loss, and validation accuracy over user-defined epoch and learning rate. Learning rate scheduling was a challenging part of the training process because there were many considerations that we had to take into such as over-shooting if the learning rate was too large and the slow-convergence rate if the learning rate is too small.

6 EXPERIMENTS - DATASET AND EXPERIMENTAL SETUP

For matching clinical trials with patient profiles[7] we are using two different data set sources to fetch clinical trials descriptions data and to fetch patient profiles data, both these sources are publicly available for analysis:

1. We used Synthea https://github.com/synthetichealth/synthea to generate synthetic patient profiles to simulate real-world patient profiles for matching with clinical trials. We also simplified the patient profiles by only fetching the necessary details relevant to our project and loading them as JSON files.

2. We fetched 50 clinical trials data from the ClinicalTrials.gov(database of clinical trials conducted in the United States) API[1] - for the following medical conditions

- Diabetes
- Cancer
- Alzheimer's disease
- Hypertension

For identifying a patient's medical symptoms[4], we are using a medical data set that provides:

- Audio file of the description of the patient's symptom
- CSV file of an overview of entire recordings

The CSV file provides insights into each patient's audio file, descriptive text translated from each audio file, and medical symptoms as labels. For each descriptive text, not the entire audio file is translated, rather only the parts relevant to categorizing medical symptoms are translated. We are using descriptive text and symptom labels to achieve supervised learning through XLNet.

For generating recommended trials using BERT we first installed necessary libraries such as nltk, matplotlib, and transformers. We then prepared synthetic patient data in JSON format and downloaded the trial data via API.

We then did exploratory data analysis for analyzing the patient details and did pre-process the data. We then loaded the pre-trained BERT model[6], computed embeddings for the cleaned trial descriptions and patient conditions, and generated the recommendations.

XLNet's setup is similar to BERT's setup in terms of the installation of necessary libraries, loading medical symptom data, preprocessing the data, and loading a pre-trained XLNet model[cite], training the additional unclassified layer with a given training data set.

Our experiments aimed to answer the research question proposed in the introduction. We did indirect evaluation of BERT model's results as outlined in the evaluation section.

To further understand the impact of different components of our method we also evaluated the performance of our models without

certain features or configurations. This allowed us to assess the significance of each component in achieving the desired results.

7 EVALUATION

7.1 Evaluation for BERT-based model

Since the model we built is a kind of recommendation system, and we did not have ground truth data to compare the recommended trials against, the challenging part was to evaluate the performance of the model directly. However, we did some indirect evaluation by analyzing the diversity of the recommendations, and the distribution of trial features (study types and enrollment sizes) within the recommended trials. By analyzing these distributions, we can assess the diversity of the recommendations and ensure that patients with varying needs and preferences have access to a variety of clinical trials.

From the below graph Figure6 we can see the recommended trials across study types. As the trials are matched with patient profiles and each patient gets the top 3 recommended trials, we can see overall how many observational trials are recommended across all the patients and how many of the interventional trials are recommended across the patients. From the graph Figure7 we can also see only a portion of patient profiles are recommended compared to the total number of enrollment sizes across the trials from figure5.

7.2 Evaluation for XLNet

We were able to directly evaluate the XLNet model's performance through training and testing the medical symptom data set provided. Given 5 epochs, average training loss and average validation loss decreased to 0.39 and 0.19 respectively. The best training accuracy achieved was 91.0%, and validation accuracy was 92.85%. With the test data set, an accuracy of 89.32% was achieved.

XLNet Model Training and Testing Result: 5 Epoch		
	Accuracy(%)	Loss
Training	91.0	0.39
Validation	92.85	0.19
Testing	89.32	-

8 DISCUSSION AND RESULTS

In this project, we learned that matching patients to clinical trials can be significantly improved using NLP techniques. We discovered that pre-trained models like BERT are powerful tools for generating embeddings that capture rich semantic information, leading to more relevant recommendations. Also, we found that the XLNet model is effective in identifying medical symptoms with high accuracy due to its generalized autoregressive structure, overcoming some limitations of the BERT model.

1. We did a recommendation and matched patients with relevant clinical trials based on their conditions and the trial descriptions. We used natural language processing techniques and a pre-trained BERT model to generate embeddings for both trial descriptions and patient conditions. These embeddings are then used to calculate similarity scores and recommend the top-k trials for each patient based on eligibility criteria, ensuring that the recommendations align with patients' needs and preferences. We utilized the power of the BERT model which was pre-trained on a large corpus of text, and using BERT for generating embeddings enabled us to capture rich semantic information from both patient conditions and trial descriptions, leading to more accurate and relevant recommendations.

2. In order to identify the patient's medical symptoms, we used supervised learning to train the XLNet model with the given medical symptom data set. We were able to achieve 89.32% testing accuracy through training. The model is a generalized autoregressive model, so the dependency on all previous tokens is heavy[8]. It overcomes the disadvantages of the BERT model, where the model attempts

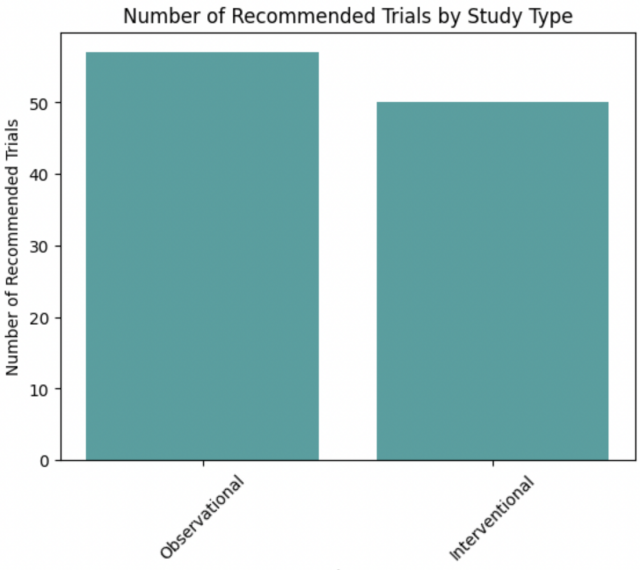

Figure 6: Distribution of type of Study across recommended trials

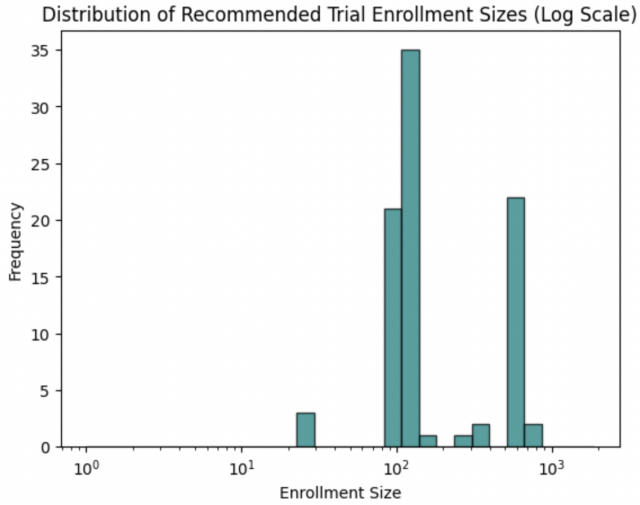

Figure 7: Enrollment size frequency for recommended trials

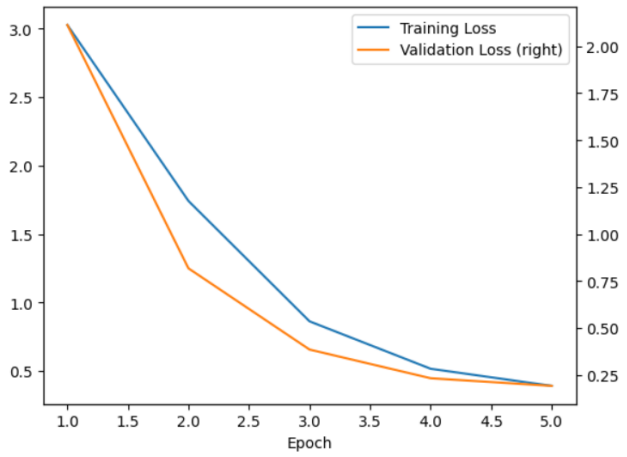

Figure 8: Training Loss and Validation Loss

to reconstruct the masked tokens independently of one another token.

8.1 Challenges

We learned the application of transformer-based models and the powerful versatility, adaptability, and robustness of the models, and how they can be adapted to any kind of Natural language task at hand.

In our proposal, we had initially planned for a comparison-based approach of using BERT and XLNet for generating recommendations. But since each model had its own shortcomings with respect to the handling of input data, we had to change our process of comparison to a sequential approach of using both the models in the medical field, using BERT for recommendation and XLNet for medical transcription identification.

The challenge was in doing a direct evaluation of the BERT model, as our task was kind of unsupervised and we didn't have ground truth data, we had to do an indirect evaluation of our model.

If we had to extend our project, we would think of merging both the BERT and XLNet model's results to generate more accurate and suitable recommendations for hospitals, doctors, and clinical trials.

9 LINK TO GITHUB REPOSITORY

https://github.com/SandhyaArumugamKarunanithy/Clinical-Trial-Matching-using-BERT-AND-XLNET

The above link contains the dataset for clinical trial matching and medical system recommendation and their respective Python files.

We also also included the final presentation in the github repository.

REFERENCES

[1] Anthony Nguyen Hamed Hassanzadeh, Sarvnaz Karimi. 2020. Matching patients to clinical trials using semantically enriched document representation. https://www.sciencedirect.com/science/article/pii/S1532046420300344
[2] Huggingface. 2020. XLNet: Generalized autoregressive pretraining for Language Understanding. https://arxiv.org/abs/1906.08237
[3] Huggingface. 2023. Transformers. https://huggingface.co/docs/transformers/v4.28.1/en/model_doc/xlnet#transformers.XLNetTokenizer
[4] Kaggle. 2019. Medical speech, transcription, and intent. https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent
[5] Abhishek Pandey Nina Arya Gwendolyn Halford Sandra F Jones Richard Forshee Mark Walderhaug Taxiarchis Botsis Kory Kreimeyer, Matthew Foster. 2027. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. https://www.sciencedirect.com/science/article/pii/S1532046417301685
[6] Taku Kudo and John Richardson. 2018. Sentencepiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. https://arxiv.org/abs/1808.06226
[7] Andrew Parchman Zhihui Luo1 Licong Cui1 Patrick Mergler Robert Lanese4 Jill S. Barnholtz-Sloan Neal J. Meropol Satya S. Sahoo1, Shiqiang Tao1 and Guo-Qiang Zhang. 2014. Trial prospector: Matching patients with cancer ... - sage journals. https://journals.sagepub.com/doi/pdf/10.4137/CIN.S19454
[8] XLNET. 2023. Using XLNet for Sentiment Classification. https://towardsdatascience.com/what-is-xlnet-and-why-it-outperforms-bert-8d8fce710335