**Problem Overview**

Primary health care is essential for most people, but making an appointment with a doctor could be exhausting and time-consuming in modern life. It would be great to have an AI healthcare chatbot to assure the anxiety before making the reservation. Image the scenario, we can make a self-diagnosis right after having some symptoms, and then decide which specialists to visit.

We attempt to predict clinical diagnosis from medical transcripts/anamnesis, although the prediction might not be accurate at all. On one side, patients can save time and money with a positive attitude; on the other side, doctors can serve more efficiently on sparing time on unreliable diagnosis.

**Data**

We will use the medical transcription sample reports and examples from [https://www.mtsamples.com](https://www.mtsamples.com/). The data contains around 5000 sample transcription reports for many specialties provided by various transcriptionists and users and are for reference purpose only. The data is also available on Kaggle (<https://www.kaggle.com/tboyle10/medicaltranscriptions>) . In our problem setting, we are interested in two aspects of the data: the sample medical transcription as features and medical speciality as labels. In the dataset, 40 types of specialties are provided as labels with an unbalanced distribution ((min,max) = (7, 1103)).

**Method**

We would like to implement multiple word embedding models to decode the medical transcripts, including the traditional count vectorizing, TF-IDF, Word2Vec, ELMo (Embeddings from Language Models) and BERT. For multi-label classification, Naive Bayes and logistic regression will be the simplest to implement, recurrent neural networks will also be a good deep learning model to boost the performance.

We intend to use the existing libraries in tensorflow or Pytorch to implement those models. We will use packages like Pandas for data preprocessing, Seaborn for visualization.

**Related Work**

[1]. Magna, Andrés Alejandro Ramos, et al. "Application of Machine Learning and Word Embeddings in the Classification of Cancer Diagnosis Using Patient Anamnesis." Ieee Access 8 (2020): 106198-106213.

<https://ieeexplore.ieee.org/ielx7/6287639/8948470/09108225.pdf>

Their work applied word-embedding representation models (word2vec’s skip-gram and BERT) methods for a recommendation system for the diagnosis of breast cancer using 160,560 medical histories of anonymous patients, based on five labels. This paper gives a detailed pipeline on how they applied BERT to their large dataset, which would be useful when we use BERT in our data set.

[2]. Sadman, Nafiz, et al. "Recommend Speciality Doctor from Health Transcription: Ensemble Machine Learning Approach."

<https://www.researchgate.net/profile/Nafiz-Sadman/publication/348335070_Recommend_Speciality_Doctor_from_Health_Transcription_Ensemble_Machine_Learning_Approach/links/5ff86223a6fdccdcb83be5a8/Recommend-Speciality-Doctor-from-Health-Transcription-Ensemble-Machine-Learning-Approach.pdf>

This paper used the same dataset from mtsamples, they ensembled five classification machine learning algorithms including Naive Bayes, K- Nearest Neighbour (KNN), Support vector machine (SVM), Random Forest classifier, and Logistic Regression, one deep learning algorithm Bidirectional Long Short Term Memory (BiLSTM). They balanced each class to 355 entries with only 4 classes in medical speciality: Surgery, Cardiovascular / Pulmonary, Consult - History and Phy and Orthopedic. This paper used classic text representations(TF-IDF Vectorizer). We would like to compare our results to this paper.

[3]. Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018).

https://arxiv.org/abs/1802.05365

This paper mainly discussed how to vectorize words through a deep learning framework. ELMo (Embeddings from Language Models

) vector assigned to a token or a word depends on current context and is actually a function of the entire sentence containing that word. So the same word can have different word vectors under different contexts. We will also use this embedding method to vectorize the transcription samples and compare its performance with other word representation models.

[4]. Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. (2019). Evaluating word embedding models: Methods and experimental results. APSIPA Transactions on Signal and Information Processing, 8, E19. doi:10.1017/ATSIP.2019.12

https://www.cambridge.org/core/journals/apsipa-transactions-on-signal-and-information-processing/article/evaluating-word-embedding-models-methods-and-experimental-results/EDF43F837150B94E71DBB36B28B85E79

This paper introduced six word embedding models, including Neural Network Language Model, Continuous-Bag-of-Words and skip-gram, Co-occurrence matrix, FastText, N-gram model, and Dictionary model, which is helpful for us to transform the medical transcripts into vectors. They also talked about the desired properties and evaluation methods of these models, it showed that different evaluators focus on different aspects of word models, it might help us to find a most suitable word embedding model.

[5]. Tanha, Jafar, et al. "Boosting methods for multi-class imbalanced data classification: an experimental review." Journal of Big Data 7.1 (2020): 1-47.

<https://link.springer.com/article/10.1186/s40537-020-00349-y>

This paper reviews the boosting methods for multi-class imbalance data classification, which is exactly the same scenario with our problem. We will apply the CatBoost and LogitBoost methods if possible in our classification task. Besides, this paper also discussed the evaluation metrics in imbalanced data domains, like Multiclass-AUC, Multiclass-MCC, and G-mean. Those implementations are also valuable in our model evaluation part.

**Evaluation**

We intend to split our dataset into a training set, a validation set, and a test set, validating the performance with several measures, including accuracy, precision, recall, F1 score, AUC-ROC curve across several different models. Since this is a multiple-class classification, we will employ the metrics like Multiclass-AUC, Multiclass-MCC, and G-mean. Those metrics of the compared approaches will be summarized in a table for illustration. Also we will plot the confusion matrix as an 40\*40 heatmap. For better visualization of the word embedding, we will use low dimensional representation like t-SNE to show the clustering structure of the data.

The baseline method would be Naive Bayer and logistic regression with traditional text representations(TF-IDF Vectorizer or One-hot Encoding). We will apply the BERT or ELMo to see if complicated word vectorization can help us improve the performance. Other than that, we will also compare the results using other classifiers like neural networks or boosting methods. Furthermore, we can have an AI healthcare chatbot application, where patients can type in their symptoms, doctors can type in the medical transcript, and it will output several corresponding diagnosis results with high possibilities.