BIS550 Project Proposal

Heyuan Huang, Gaoqianxue Liu, Ji Qi, Yingxue Pan

April 6, 2023

1 Background

Patient-provider communication (PPC) refers to the analysis and interpretation of language used by patients and healthcare providers in the context of a clinical encounter. This type of NLP task involves the application of computational techniques to extract meaningful information from the textual or spoken communication between patients and healthcare providers. By applying NLP techniques to patient-provider communication, researchers and clinicians can gain insights into the quality of the interaction, identify areas for improvement, and develop tools to support clinical decision-making and improve patient outcomes.

2 Research Gaps We Plan to Address

In this project, we aim to automate the process of identifying communication patterns between patients and healthcare providers, which is crucial for effective and prompt Patient-Provider Communication (PPC). Rapid PPC is essential for enhancing population health management and achieving patient-centered outcomes. By leveraging machine learning and natural language processing, we seek to extract features and classify from patient-provider interactions, including symptoms, adverse events, medications, emotions, and expressions of empathy. This approach will allow for closer monitoring of patients' adherence and faster response to changes in their health status.

3 Solution to address gaps

- Use sciscpacy models to detect medical entities in our text. ScispaCy is a Python package containing spaCy models for processing biomedical, scientific or clinical text.
- Perform feature extraction using TfidfVectorizer to generate tf-idf features based on the PPC. In information retrieval, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
- Apply domain knowledge to reduce the categories. E.g., the category "surgery" can be a superset that contains surgeries from medical specialties like "cardiology" or "neurology". We may simply remove "surgery" to improve model performance.
- Do Principal Component Analysis (PCA) to reduce the dimensionality of features.
- Try penalized logistic regression, SVM, decision tree, and random forest to train models that predict the labels based on the PPC transcription.
- Use SMOTE(Synthetic Minority Over-sampling Technique) to generate more samples from minor class to solve the data imbalance problem.

4 Dataset

The dataset includes a total of 2358 transcription samples. The transcription samples are divided into 40 categories based on medical specialty, such as allergy, autopsy and etc. For each transcription, there is a short description and sample keywords.

5 Computational methods

We plan to use penalized logistic regression, SVM, decision tree, and random forest to finish the classification tasks. Logistic regression is a statistical method used to analyze and model the relationship between a dependent variable and one or more independent variables, while SVM, or support vector machine, is a supervised learning algorithm used for classification and regression analysis that separates data points into different classes based on a hyperplane, and decision tree and random forest are ensemble learning algorithms that use multiple decision trees to classify data points and make predictions.

6 Roles and contributions of team members

- Project manager: Heyuan Huang
  - Coordinating team meetings and delegating tasks
  - Monitoring progress
  - Communicating with the professor and TA
- Researchers: Gaoqianxue Liu, Yingxue Pan, Ji Qi
  - Data cleaning and preprocessing
  - Conduct feature Extraction and dimensionality reduction
  - Construct machine learning models and hyperparameter tuning
  - Testing and model comparison

7 Timeline

- March 20: Data preprocessed and cleaned.
- March 31: Plan in this proposal finished.
- April 7: Model performance assessed and method framework improved.
- April 15: Final manuscript and slides finished.
- April 25: Final project presented in class.