ABSTRACT

Hospitals capture large amounts of text data for each admitted patient. This data gets stored as part of the patient's progress notes. Many areas of the hospital submit these notes. Primarily used for day-to-day communication between the patient's care team, they may be beneficial for other purposes. This project explored whether progress notes could be used to support the diagnostic decision. Are there text data hidden in the progress notes that can help us find commonalities and distinctions between diagnoses? This study analyzed progress notes on a subset of the 54,000 admissions available in the Medical Information Mart for Intensive Care, MIMIC-III, an MIT-curated healthcare database.

While patients are in the hospital, especially before a diagnosis, a patient's health and behavior are closely monitored and captured. Much of this data is captured in patient progress notes entered by all staff members participating in the patient's care. This data is valuable input during the diagnostic assessment, but the volume of data is both a benefit and a curse. The volume generates a 'cannot see the forest for the trees' problem. The pertinent information is hidden in a forest of data. As other researchers have noted, new electronic medical record systems exacerbate this problem, since a nurse or physician often creates a new note by copying an old note and adding some additional comments. This adds more 'less-useful' trees in the forest, making the forest larger and the important trees harder to find.

This study examined the ability of two text analysis approaches to find pertinent information to determine the likely diagnosis. The predicted diagnosis code signifies that the patient's notes fit a pattern to other patients with a similar diagnosis code. It can be used by the physician, with all other available data, to make the final diagnostic decision. If the assigned diagnosis code is drastically different, it can also be used in the way Stockwell referred to as

‘trigger data’ that drives a more in-depth review. Since early diagnosis is important in any medical intervention, the goal is to use only the first three days of notes so that the prediction can be used early enough in the patient’s in-patient stay to benefit the physician’s diagnostic decision.

The main approaches used to analyze the notes were: first, Term Frequency-Inverse Document Frequency (TF-IDF) matrix used as input to a Random Forest Classifier to predict the diagnosis, and secondly, a fine-tuned Bidirectional Encoder Representations From Transformers (BERT) Large Language Model to predict the diagnosis. Hyperparameter tuning was completed using cross-validation and response surface methodology (RSM). The models were executed standalone and as part of an ensemble model. The developed models were trained and tested using two groups of four diagnosis codes: one set that was very distinct from each other, i.e., affected different body systems, and another that was similar. i.e., all related to the heart. Being able to identify diagnosis codes in these two situations is very important. The full combination of model characteristics and diagnosis code groups resulted in the development of ten fitted and tested models.

In analyzing the distinct diagnosis codes, an ensemble model using average probability technique combined the results from the RF-TFID model and a BERT fine-tuned model using RSM to develop hyperparameters performed the best. Its accuracy was 86.7% with an MCC of 0.8222. For the similar diagnosis code group, an ensemble approach with the same characteristics as the distinct one also outperformed the other with an accuracy of 76.7% and an MCC of 0.6889.

There is a valid concern in the field of artificial intelligence that models may contain undetected biases. Poor data selection can be one cause of this problem. MIT’s MIMIC-III

database of ICU data was the source of our research data. It was noted during exploratory data analysis that the patient population for this data is predominantly white with minority representation percentages for our selected diagnosis codes smaller than the estimated US minority percentages. Therefore, data bias was a concern here. To address this concern a second set of validation tests were executed on the distinct diagnosis models using 100 minority patient admissions as input. These tests were run as a group of all minority admissions, as well as a set of just black admissions. The results of those tests show that the best ensemble model on the earlier tests identified the diagnosis code for the minority patients 82%, accurately with an MCC > .7600. However, a model that did not do well in the earlier tests, Random Forest using TF-IDF, actually performed better on this population than all other tests. Its accuracy was > 88.6% with an MCC > .8476. This demonstrated that these models were effective regardless of patient ethnicity, and that better results can be achieved if you consider ethnicity in your model choice.