Predicting Medical Diagnoses from Drug-Related Google Search Text using Natural Language Processing (NLP) on MIMIC-IV Dataset

Brian Lewis
Northwestern University, MS in Analytics
brian.lewis@u.northwestern.edu

Abstract

Linking prescription drug information with appropriate medical diagnoses can be a time-intensive process with the frequent changes of National Drug Code (NDC) identification numbers. In this paper, I report the performance of a natural language processing model that can map text from Google's "Custom Search" API about prescription drugs associated with medical diagnosis codes. This model can predict the primary diagnosis from unstructured entries about a particular drug and its medical indications. Given the many-to-many dynamics of multiple prescriptions and multiple diagnoses, my best performing model predicts the top-50 ICD-9 and ICD-10 medical codes with a Micro-F1 Score of 13.0% , an improvement over a naïve classifier for 50 outcomes.

1 Introduction

Electronic health records (EHR) contain valuable patient-level information regarding medical history, prescription drug use, and medical diagnoses for both inpatient and outpatient settings. Understanding the relationship between a patient's prescription drug use and medical diagnoses can aid physicians, insurance companies, and pharmacists in 1) validating that patients are receiving the correct prescriptions and diagnoses given one of the two pieces of information and 2) gaining a better knowledge of a medical patient's probable diagnoses given their current prescription medication regimen in the absence of other prior medical EHR data.

Natural Language Processing (NLP) is not a new set of tools within the healthcare domain, but the author of this paper is unaware of its use to attempt predicting ICD-9 and ICD-10 medical diagnoses using unstructured text about prescription drug data. Doing so is an inherently complex problem given the myriad uses of a particular drug with a variety of potential medical diagnoses (e.g. a doctor may recommend using intravenous fluids for patients admitted to a hospital for hundreds of different reasons). This paper seeks to address one approach among many possible approaches to predict medical diagnoses from unstructured text data about prescription drugs using the MIMIC-IV dataset.

This paper will first review related academic work, describe the MIMIC-IV dataset and necessary preprocessing steps, walk through the methods used to build the NLP model itself, present the results of the model, and discuss the interpretation of the results as well as limitations and future work related to this approach.

2 Related Work

Some form of automatic ICD diagnosis coding has been around for decades, initially using discharge summary notes to classify a diagnosis (Larkey and Croft, 1995). More recently, researchers have continued to build more accurate models classifying a patient's diagnosis based on clinical notes using more powerful models such as ULMFiT (Nuthakki et al, 2019), Hierarchical Label-wise Attention Networks (Dong et al, 2021), and BERT (Heo et al, 2021). Each of these prior papers relied upon data found in the MIMIC-III dataset, which is very similar to the MIMIC-IV dataset used in this paper and which will be discussed in greater depth below. Another paper (Huang, et. al; 2019) carries out a type of meta-analysis of several techniques and compares them against one another in the MIMIC-III dataset.

Nuthakki et al employed the ULMFiT architecture of a neural network to use on unstructured text from several sources, including patient illness history, symptoms at time of admission, and other clinical notes to predict medical diagnosis. The authors of this paper also acknowledge the immense complexity of patients receiving multiple, and often seemingly unordered, medical diagnoses. Nuthakki et al reduce this complexity by filtering the data to consider only the first and (and intended to be most important) diagnosis as a result.

Dong et al focuses on clinical notes within the MIMIC-III dataset but utilizes Hierarchical Label-wise Attention Networks (HLAN) to quantify the importance of words and sentences in obtaining medical diagnosis codes. These authors also benchmarked the results of their HLAN method against other network-based models such as CNN and RNN and found that their HLAN technique yielded higher accuracy than CNN-based models.

Heo et al followed a similar approach to Nuthakki et al, but used BERT in order to increase the utility of word embeddings from clinical notes in predicting medical diagnoses. Heo et al narrowed down their predictions to the top-50 most frequent medical diagnoses for simplicity as Nuthakki et al did.

Huang et al reviewed the efficacy of more traditional machine learning (ML) methods in the same tasks as listed above, and then compared those results against the more cutting-edge deep learning models. They found that a logistic regression model obtained the highest F1-score when looking at the top-50 medical diagnosis codes. Huang et al found that only when looking at a smaller subgroup of diagnoses did deep learning models perform better than traditional ML methods.

3 Dataset and Preprocessing

MIMIC-IV is a large dataset relating to patients admitted to the Beth Israel Deaconess Medical Center between 2008 and 2019. (MIMIC-IV builds off of the existing MIMIC-III dataset, used in the above-referenced papers, and includes the same types of data from same facility. MIMIC-III is an older version of MIMIC-IV and only includes data between 2001 and 2012.)

Obtaining access to MIMIC-IV is somewhat complicated. It requires completing a data compliance course and receiving credentials from researchers at the Massachusetts Institute of Technology. It took a few days to complete the process and finish signing the various data use agreements.

MIMIC-IV includes a vast set of available tables, but for the purposes of this paper, only three primary tables were considered: "diagnoses_icd", "d_icd_diagnoses", and "prescriptions". This raw data provided hundreds of thousands of hospital admissions with millions of prescription drugs associated with the admissions and millions of medical diagnoses.

As Nuthakki et al pointed out, each patient's admission in the MIMIC-IV had multiple diagnoses, creating a "many-to-many" mapping problem when joining patient prescription data. Following the documentation of MIMIC-IV and Nuthakki et al, the data were filtered to consider only the first (most important) diagnosis for each hospital admission. The data were further filtered by narrowing down observations to those which included a diagnosis within the set of the top-50 most frequent diagnoses inside the MIMIC-IV dataset, as had been done by every other paper referenced previously.

The "prescriptions.csv" data also required additional pre-processing. Many of the National Drug Code (NDC) ID numbers listed with prescription data observations did not match existing NDCs in the OpenFDA Drug API as initially planned. Without the ability to match these NDCs together, the 'drug indications' data for the prescriptions observations was unavailable.

In order to proceed with the project using a creative solution, I created a paid Google Cloud API account and collected JSON text responses of the first-page results for the following search query for each unique drug name in the filtered prescriptions dataset: "What is <drug name> used for?". These JSON text responses from Google's "Custom Search" API were further cleaned and processed, becoming the unstructured text associated with each prescription. This unstructured text for each prescription was then linked back to the medical diagnosis dataset above with patient and admission ID codes.

The last obstacle related to the dataset and preprocessing had to do with class imbalance. Even within the selected top-50 diagnoses, a handful of diagnoses accounted for a very high percentage of the total observations. Given the repetitive nature of the unstructured drug text and the abundance of total observations, both random down-sampling and up-sampling methods were employed while running modeling experiments (see below) in order to achieve an equitable balance among all 50 diagnosis classes and improve the power of the final classification model used in this paper.

Once these preprocessing steps had been taken, the resulting unstructured text for each drug was tokenized and saved with its associated primary diagnosis for use in the modeling portions discussed later in the paper. The final dataset used for modeling was approximately 3.3 GB in size and consisted of relatively low-grade, unstructured, and repetitive text as well as 50 medical diagnoses that made up the categorical classes.

4 Method

4.1 Deep Learning

Seeing the success of previous deep learning models in this domain made a BERT model a natural starting point for a preliminary model. Initial attempts were made to utilize BERT and neural networks on a subsample of the data to create a multi-label classification model. Unfortunately, even with class balancing strategies, these led to the model predicting the same 5 labels across nearly all observations. This result, combined with the extremely lengthy training times and computational expense meant that deep learning approaches were ultimately abandoned for more traditional ML models in the hopes of recreating a similar outcome to previous findings with logistic regression (Huang et al, 2019).

4.2 Traditional ML (Logistic Regression)

Logistic regression was chosen as the next plausible model given the results found by others (Huang et al, 2019). Preprocessed data were read in, split into train and test sets on an 80-20% basis (respectively), and then two sets of experiments were run. The first set involved random over-sampling of minority class training data until all 50 classes were balanced, the second set involved random under-sampling of the majority classes in the training data until all 50 classes were balanced. As can be seen in Table 1 below, the random over-sampling strategy yielded better results on the test set.

Following class balancing, data were vectorized using a TD-IDF vectorizer and then run through various logistic regression classifiers on the 50 balanced classes. Hyperparameters were altered throughout the varying experiments. These hyperparameters included: bag of words representations (e.g. 1gram, 2gram, and 1gram+2gram), TD-IDF maximum threshold, and regularization norms. A variety of experimental metrics were measured which will be discussed below.

5 Results

The final results can be seen below in Table 1. Additional experiments that are not reported were conducted with smaller subgroups of the full dataset in order to save time and streamline the experimentation process.

N-grams	TD-IDF Max	Norm	Micro F1
1	2	L1	13.00%
1+2	2	L1	11.83%
2	2	L1	13.00%
1	2	L2	12.13%
1+2	2	L2	11.27%
2	2	L2	12.15%

Table 1: Results of Experiments.

6 Discussion

6.1 Interpretation

Among the models and experiments run, the best performing model across all dimensions (including precision, recall, and traditional F1-score) was the unigram, L1-norm logistic regression model. This model, run across all of the observations described in Section 3, resulted in a Micro F1-Score of 13.0%. The full classification report can be seen on the next page in Table 2.

Even with attempts at class balancing, there are still large problems with the classifier. As seen in Table 2, many of the top-50 diagnoses are never assigned any classifications. The most frequent diagnoses are still the default for many classified observations.

6.2 Limitations and Future Work

The scores presented in Table 1 are much lower than the F1-Scores included in the related academic work. But there are some distinct are some obvious problems with this dataset relative to the data used in the related work. In the academic literature cited above, clinical notes were used. Each of these is also inherently distinct, and written in coherent speech.

In the present example, incoherent (non-speech) text data was gathered from Google's "Custom Search" API for each drug. Given that the same drugs were used for different diagnoses, these were not distinct text entries. Further, unlike clinical notes, drug descriptions may indicate something entirely different from how a particular physician utilizes the drug in a specific medical scenario. Given these complexities, it is unsurprising, then, that the results are as low as they are.

Proper future work would involve obtaining 1) cleaner drug indications, 2) additional text data from clinical notes, and 3) a more balanced dataset in order to build a better classifier for this topic.

diagnosis	precision	recall	f1-score	support
10_A419	0.12	0.17	0.14	23565
10_F10129	0	0.68	0	166
10_F329	0	0	0	990
10_I110	0	0	0	7014
10_I130	0.24	0.03	0.06	11321
10_I214	0.96	0.01	0.01	15406
10_J189	0	0	0	7204
10_N179	0	0	0	7992
10_N390	0	0	0	5741
10_R0789	0	0	0	2460
10_R079	0	0	0	987
10_Z3800	0.18	0.44	0.25	11090
10_Z3801	0	0	0	6624
10_Z5111	0.2	0.08	0.12	11705
9_0389	0.15	0.14	0.14	22297
9_27651	0	0	0	3673
9_2989	0	0	0	1086
9_30500	0	0	0	195
9_311	0	0	0	900
9_41071	0.19	0.02	0.04	18277
9_41401	0.15	0.65	0.25	36279
9_41519	0	0	0	5022
9_4241	0	0	0	13622
9_42731	0.17	0.08	0.11	11380
9_42823	0	0	0	9964
9_42833	0	0	0	11688
9_431	0	0	0	6324
9_43491	0	0	0	5044
9_486	0.15	0.12	0.13	16961
9_49121	0	0	0	5710
9_5409	0	0	0	2728
9_5609	0	0	0	4236
9_56211	0	0	0	5802
9_5770	0.13	0.03	0.05	11275
9_5849	0.08	0.03	0.04	14944
9_5990	0.15	0.02	0.03	11512
9_64511	0.54	0.04	0.08	3294
9_65421	0	0	0	3690
9_6826	0	0	0	7355
9_71536	0	0	0	6258
9_72210	0	0	0	4641
9_7802	0	0	0	5876
9_78097	0	0	0	1251
9_78650	0	0	0	3016
9_78659	0	0	0	9668
9_99859	0	0	0	10345
9_V3000	0.44	0.46	0.45	18571
9_V3001	0.26	0.18	0.22	11894
9_V3101	0	0	0	3477
9_V5811	0.23	0.27	0.25	15655
accuracy			0.13	436175
macro avg	0.09	0.07	0.05	436175
weighted avg	0.14	0.13	0.10	436175

Table 2: Classification Report of Best Model

7 Source Code

Final source code can be found here: https://github.com/MSIA/btl1613_msia_text_analytics_2021/tree/project/project.

References

Hang Dong, Victor Suarez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of Biomedical Informatics, volume 116, p. 103728.* https://doi.org/10.1016/j.jbi.2021.103728.

AL Goldberger, LAN Amaral, L Glass, JM Hausdorff, PCh Ivanov, RG Mark, JE Mietus, GB Moody, C-K Peng, HE Stanley. 2000. PysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resourced for Complex Physiologic Signals. *Circulation, volume 101(23):e215-e220.* http://circ.ahajournals.org/content/101/23/e215.

Tak-Sung Heo, Yongmin Yoo, Yeongjoon Park, Byeong-Cheol Jo, and Kyungsun Kim. 2021. Medical Code Prediction from Discharge Summary: Document to Sequence BERT using Sequence Attention. Computer Science > Artificial Intelligence submission to arXiv.org. https://arxiv.org/abs/2106.07932

Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. 2019. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Computer Methods and Programs in Biomedicine, Volume 177, pp. 141-153.* https://doi.org/10.1016/j.cmpb.2019.05.024.

A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark. 2021. MIMIC-IV (version 1.0). *PhysioNet*. https://doi.org/10.13026/s6n6-xd98.

Leah S. Larkey and W. Bruce Croft. 1995. Automatic Assignment of ICD9 Codes to Discharge Summaries. Technical Report, University of Massachusetts at Amherst, Amherst, MA. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.816&rep=rep1&type=pdf

Siddhartha Nuthakki, Sunil Neela, Judy W. Gichoya, Saptarshi Purkayastha. 2019. Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks. Computer Science > Computation and Language submission to arXiv.org. https://arxiv.org/abs/1912.12397.