

Using LLMs for Cancer Diagnosis from Clinical Notes

ISAAC HANDS, University of Kentucky, USA

1 BACKGROUND AND INTRODUCTION

Cancer is a devastating illness, the second leading cause of death in the U.S [?]. One of the best ways to improve survival of this disease is to diagnose it as early as possible[?]. The way cancer is diagnosed in 90% of cancer cases[?] is through a biopsy of tissue, which is then observed under a microscope by a pathologist, leading to a narrative description of the cancer cells in a pathology report. These reports consist of clinical narrative text dictated by a pathologist, stored in a patient's electronic medical record. The characterization of a cancer diagnosis occurs when a physician interprets a pathology report for communication to the patient and determines a treatment plan, along with data abstraction of the codes associated with the cancer diagnosis by specialized medical coders. This entire process of determining whether a patient has cancer and what type of cancer it is, relies on interpretation of narrative text. In fact, narrative unstructured text makes up about 80% of all healthcare data[?], making it difficult to analyze systematically.

Processing, analyzing, and interpreting unstructured text can be time consuming and cause delays in life-saving treatment for cancer patients, but may be an area where machine learning classifiers could provide efficiencies. In order to build classifiers, however, labeled training data needs to be readily available. When the data required is clinical text containing patient identifiers, it is often very difficult to access due to privacy concerns and lack of access to medical record systems. Moreover, clinical text uses a specialized vocabulary, which may prove problematic for natural language processing methods that have been developed around standard English vocabulary. In this report, we attempt to overcome these difficulties by utilizing a large language model (LLM) pre-trained on biomedical text, fine-tuned with clinical text documents that have a cancer diagnosis label. We will investigate both BERT[?] and GPT[?]-based transformer[?] models, which have been shown to perform well on machine learning tasks in the biomedical realm[?] [?].

2 MATERIALS AND METHODS

We identified a dataset of clinical notes labeled with cancer diagnoses, along with BERT and GPT-type LLM models that could be fine tuned on standard hardware. The clinical note dataset we utilized, MIMIC-IV-Note[?], required a request for credentialed access for academic use from the authors, which was granted within one day. Even though the data was narrative text from actual hospital patients in a US, it was de-identified through an automated NER task[?] before being made available, so privacy and security was not a concern. We also identified four transformer models, two BERT and two GPT, from the HuggingFace open source library, that were well characterized and able to be trained on a M1 Macbook Pro. Our programming for dataset preparation and model fine-tuning was done in Python 3.9 utilizing the PyTorch library[?]. The source code used to process data and fine-tune the models is published in GitHub[?].

2.1 Data Description and Processing

The MIMIC-IV-Note dataset consists of 331,794 discharge summaries and over 2 million free-text radiology reports representing clinical notes on more than 300,000 patients at the Beth Israel Deaconess Medical Center in Boston, MA. For this study, we focused on the discharge summaries due to our limited computational power available and the volume

Author's address: Isaac Hands, isaac.hands@uky.edu, University of Kentucky, USA.

of the data. The data elements for each clinical note in the MIMIC-IV-Note dataset consisted of the actual narrative text of the note, three ID fields, and two timestamps[?]. In order to get labels for each note, the MIMIC-IV dataset[?] was used (note the difference between the names of the two datasets: MIMIC-IV-Note and MIMIC-IV). The MIMIC-IV dataset contains ICD-9 and ICD-10[?] labels for many of the clinical notes in the MIMIC-IV-Note dataset, linked by the three ID fields. Older notes had the older ICD-9 codes while newer notes used ICD-10. Each clinical note was labeled with multiple codes, assigned by medical billing coders when the patient was discharged from the hospital.

The first step in processing the MIMIC-IV dataset was to filter out all labels that were not ICD-10 codes. We decided to focus exclusively on ICD-10 since it is a newer coding standard and to reduce the size of the fine-tuning dataset. Next, we filtered the MIMIC-IV-Note dataset to only include clinical notes that have ICD-10 labels and then joined the clinical notes with a list of their complete ICD-10 labels using the three ID fields. To simplify our machine learning task, we developed a binary cancer diagnosis classification scheme by labeling each clinical note as 'no cancer' (0) or 'cancer' (1) based on whether any of the ICD-10 labels for a note indicated a cancer code. In the ICD-10 coding vocabulary, all codes related to malignant cancers begin with the capital letter 'C', and no other ICD-10 codes have the capital letter 'C', so this binary labeling step was reduced to simply looking for the 'C' character among all of the ICD-10 codes. After the binary labels were attached to each note, we removed all data fields except the text and the label.

The final step in processing the data set was to create randomized, split, balanced datasets, of varying sizes for training and development. Since most patients in the hospital system did not have a cancer diagnosis, we created balanced datasets where the number of notes labeled as cancer (1) matched the number of notes without cancer (0). We also wanted to have small, medium, and larger dataset sizes in order to quickly and progressively run the source code during development. We split the datasets into 70% training and 30% testing sizes, shuffled randomly in a balanced manner across both labels. The final balanced dataset sizes were: 100, 1000, 10000, and 22601.

2.2 Models Utilized

We selected four models from the HuggingFace open source model library for our experiments:

- stanford-crfm/BioMedLM: a GPT based language model trained only on biomedical abstracts and papers from The Pile
- gpt2-xl: a pretrained language model based on GPT-2. The model is trained on the English language using causal language modeling.
- dmis-lab/biobert-base-cased-v1.2: a BERT based pretrained language model trained on PubMed abstracts and PMC full-text articles
- emilyalsentzer/Bio_ClinicalBERT: initialized from BioBert and trained on all MIMIC-III notes

Two of the models tested were GPT and two were BERT-based, giving us a mix of encoder and decoder models from the transformer architecture.

3 RESULTS

4 DISCUSSION