Using LLMs for Cancer Diagnosis from Clinical Notes

ISAAC HANDS, University of Kentucky, USA
ANCHIT BHATTACHARYA, University of Kentucky, USA

1 BACKGROUND AND INTRODUCTION

Cancer is a devastating illness, the second leading cause of death in the U.S [2]. One of the best ways to improve survival of this disease is to diagnose it as early as possible[4]. The way cancer is diagnosed in 90% of cancer cases[3]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[10], 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[5] and GPT[14]-based transformer[15] models, which have been shown to perform well on machine learning tasks in the biomedical realm[16][11].

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[7], 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[8] 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[13]. The source code used to process data and fine-tune the models is published in GitHub[6].

Authors' addresses: Isaac Hands, isaac.hands@uky.edu, University of Kentucky, USA; Anchit Bhattacharya, abh240@uky.edu, University of Kentucky, USA.

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 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[1]. In order to get labels for each note, the MIMIC-IV dataset[9] 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[12] 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. We chose these models because they were well cited in the literature and the largest models from HuggingFace that our computational platform could train and test within our time constraints.

3 RESULTS

We trained our four models on dataset of size 1000. We divided the data using a 70/30 train test split, where we train our data on 700 samples and test it on 300 samples. We report the accuracy, precision, recall and F1 scores on our test data after training the model for 10 epochs.

Model	Parameters	Style	Accuracy	Precision	Recall	F1
stanford-crfm/BioMedLM	2.7B	GPT	0.833	0.816	0.888	0.850
gpt2-xl	1.5B	GPT	0.847	0.875	0.831	0.853
dmis-lab/biobert-base-cased-v1.2	340M	BERT	0.863	0.922	0.813	**0.864**
emilyalsentzer/BioClinicalBERT	340M	BERT	0.850	0.857	0.863	0.860

4 DISCUSSION

Based on our results we found out that the dmis-lab/biobert-base-cased-v1.2 produced the best F1 score followed by the Bio_ClinicalBERT model and outperformed the other 2 models used in this experiment. Surprisingly, the results produced by the BioMedLM was almost similar to the GPT2-XL model although BioMedLM was pretrained on bioclinical text whereas GPT2-XL was trained on general english text. For future work, we would like to explore our model on our largest dataset size of 22601, which we couldn't because of limited resources and massively large training time of the models. We would also like to explore better hyperparameter tuning on the existing experiments to see if the results can be further improved. In terms of newer methods, we would like to explore zero shot learning with prompt engineering on the GPT based models.

REFERENCES

[1] [n. d.]. Discharge Note Data Schema. https://mimic.mit.edu/docs/iv/modules/note/discharge/

[2] CDC1 2023. An Update on Cancer Deaths in the United States. Retrieved April 29, 2023 from https://www.cdc.gov/cancer/dcpc/research/update-on-cancer-deaths/index.htm

[3] CDC2 2023. How CDC Speeds Up Cancer Data Reporting. Retrieved April 29, 2023 from https://www.cdc.gov/cancer/dcpc/research/articles/cdc-cancer-reporting.htm

[4] CRUK 2023. Why is early cancer diagnosis important? Retrieved April 29, 2023 from https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Isaac Hands. 2023. CS685 Final Project. https://github.com/isaackcr/CS685FinalProject

[7] AEW Johnson et al. 2022. MIMIC-IV-Note: Deidentified free-text clinical notes (latest version). *PhysioNet https://doi. org/10.13026/1cjn-2370* (2022).

[8] Alistair EW Johnson, Lucas Bulgarelli, and Tom J Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. 214–221.

[9] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-wei H Lehman, et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data* 10, 1 (2023), 1.

[10] Hyoun-Joong Kong. 2019. Managing Unstructured Big Data in Healthcare System. *Healthc Inform Res* 25, 1 (Jan 2019), 1–2. https://doi.org/10.4258/hir.2019.25.1.1

[11] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* 23, 6 (09 2022). https://doi.org/10.1093/bib/bbac409 arXiv:https://academic.oup.com/bib/article-pdf/23/6/bbac409/47144271/bbac409.pdf bbac409.

[12] World Health Organization et al. [n. d.]. ICD-10, International statistical classification of diseases and related health problems. 2007. *World Health Organization* ([n. d.]).

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[14] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[16] Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2020. Pre-training technique to localize medical bert and enhance biomedical bert. *arXiv preprint arXiv:2005.07202* (2020).