Extracting Social Determinants of Health from Electronic Health Records Using Natural Learning Processing

Viraj Mehta, Stanley Yang, Jeremy Tian

1. Introduction

Currently, meaningful clinical data is lost within unstructured sequential data in electronic health records (EHRs). We propose training natural language processing (NLP) models on EHRs containing doctors' notes of their diagnosis and patients' conversations to uncover hidden word associations with social determinants of health (SDoH). For example, clinical notes may hint at food insecurity in patients without explicitly stating "food insecure." Social determinants of health, including economic stability, food insecurity, homelessness, access to healthcare, and education, are important factors in understanding a patient's long-term health contexts and in reducing discrepancies between healthcare and its consequences in the real world. We may begin by focusing on a single social determinant and expand to multiple determinants after.

We propose using in-context learning for this problem. Using EHRs, we ask a simple question for each patient's EHR notes - does this person have x SDoH? We want to do a simple classification task before we move onto looking for associations with this SDoH and factors like vitals and possibly predictions for prognosis of disease.

2. Background

In recent years, the use of natural language processing for medical applications, particularly with regards to EHRs, has grown. For this application in particular, Patra et al. conducted a survey of various NLP methods for extracting social determinants of health from clinical data. In particular, their approach relied on manually developing a lexicon for each SDoH category with the help of domain experts. Though the creation of these lexicons allowed for a thorough understanding of keyword associations, it was also time-consuming and resource-inefficient, as opposed to an approach involving a pretrained model finetuned on clinical data or a lexicon creation using supervised or semisupervised methods. Additionally, Patra et al. do not focus on food insecurity as an SDoH, even though food insecurity has a very high correlation with a variety of health outcomes (i.e. 20% increased risk of hypertension, 3 times higher odds of having anemia, increased risk of birth defects, cognitive problems, and anxiety).

Though Patra et al. is the main significant contribution of note to applying NLP towards SDoH specifically, natural language processing has been applied in other contexts towards extracting clinical data from EHRs. For example, Khurshid et al. applied NLP to EHRs in order to recover vital signs from unstructured clinical notes. Khurshid et al. evaluated the use of Bidirectional Encoder Representations from Transformers (BERT), but more specifically Bio + Discharge Summary BERT, a contextual word embedding model pretrained on a large corpus of English text as well as biomedical text. After training for 2 epochs, the BERT model demonstrated high accuracy (96% for weight, 100% for height and blood pressure). Thus, Khurshid et al. validate the use of the BERT model as successful in contextualizing word embeddings when applied to unstructured clinical data.

3. Dataset

We harnessed data in the form of electronic health records (EHRs) from the Stanford Medicine Research Data Repository (STARR). This data has been collected from the Adult Hospital and the Lucile Packard Children's Hospital under IRB approval. The data includes information on names, dates, medical record numbers, account numbers, demographic identifiers, lab/test results, diagnosis or procedure codes, clinical narratives/reports, prescriptions or medications, and imaging reports.

Our dataset has been manually labeled with a true or false indication for food insecurity, the social determinant of health we've chosen to focus on in this project. The dataset was labeled with the help of Akshay Swaminathan in the Stanford School of Medicine. In terms of the breakdown of the demographics of our dataset, we have a total of 259 clinical patient notes. Of those 259, 120 were male patients, 131 were female, and 8 were unclassified/other. Additionally, 84 patient notes were labeled true for the existence of food insecurity, while 175 were labeled false. One potential limitation of this dataset is that it is very small (only 259 total patients); however, this leads to our choice of methodology for this project involving in-context learning, an emergent property of LMs that allows for very fast training on a low amount of data.

**Dataset breakdown visualizations in appendix*

4. Approach

4.1 Methods: Baseline

Currently, there are various powerful natural language processing models such as GPT-2, GPT-3, XLNet, T5, ELECTRA, BERT, and other modifications of BERT. However, unlike other models before it, BERT does not use unidirectional learning to perform word predictions. Technically, BERT is classified as using bidirectional learning; however, it would be more accurate to describe BERT as a non-directional learning model. With unidirectional learning models, typically the previous words in the sentence are taken into account as context to predict the next word. With BERT though, it does not predict masked words in sentences using exclusively past words, but instead reads the whole sample text at once and uses that as context to inform its decision of what it thinks the next word is. BERT does this by applying masked language modeling, MLM, which replaces a word with a mask and forces BERT to use the context from words before it and after it to predict the current masked word. Between sentences, BERT is also pre-trained through a task called Next Sentence Prediction (NSP) in order to learn relationships between different sentences in a text. In NSP, sentences are paired with correct follow-up sentences and incorrect follow-up sentences, driving BERT to improve its understanding of meaning sentences by asking it to predict if the pairing is correct or incorrect. Through these two exercises and training through massive corpuses of data, BERT is able to achieve high performance, making it one of the most robust NLP models. This is one of the primary reasons we decided to use BERT as our baseline model in our task of predicting social determinants of health given electronic health records.

As such, we decided to establish a baseline of performance prior to attempting in-context learning to understand how a pre-trained BERT model would perform when simply fine-tuned to our data and hyperparameters. To find the best hyperparameters for our model, we used a grid search to methodically evaluate each model based on varied hyperparameters. We first split our data into a training, validation, and test set to evaluate the model on our data. We then tokenized and encoded the sentence sequences

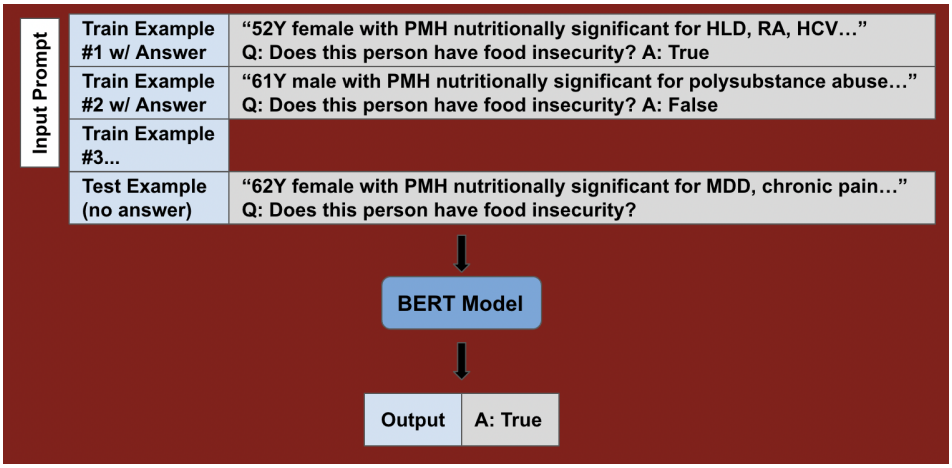
using the BERT tokenizer, padding to a maximum sentence length of 100, and truncating past the maximum for extended sequences. In our forward propagation, we used a leaky relu activation and a softmax activation function but also froze the preexisting parameters while running our model. Finally, we used an Adam optimizer with a learning rate of 0.00001 and a binary cross entropy loss function. After running BERT on our electronic health record data without in context learning, we established that the model achieved an accuracy of about 65%. The precision for false predictions was 0.69 with a recall of 0.86 and an F1-score of 0.77. On the other hand, the precision for true predictions was 0.45 with a recall of 0.24 and an F1-score of 0.31. This acted as our baseline model to compare our in-context learning integration model with.

4.2 Methods: In-Context Learning

In order to improve upon our baseline of a pre-trained BERT model, we decided to incorporate an emergent property of language models that has not yet been fully explored: in-context learning. Though currently BERT is a very robust language model, it can only form associations between different words. Thus, if we ask it to determine if a certain patient has food insecurity, it will only be able to give us an answer based on word associations. In contrast, by performing in-context learning, we can give the model a question: “does this patient have food insecurity?”, and the model will begin to learn what food insecurity actually means based on the clinical notes instead of just based on word associations.

In current literature, in-context learning has primarily been applied to the GPT-3 language model. As an autoregressive model, GPT-3 performs text generation by generating tokens one-by-one, which are then later transformed into strings of text. However, we propose using in-context learning with a new language model, BERT, which generates contextual embedding representations of each token. BERT is another example of a large language model trained using massive amounts of data. In-context learning is an emergent property specific to large language models, as LMs pretrained on massive corpuses of text allow the model to learn latent concepts. Using in-context learning is especially important for industries where there is not an abundant amount of labeled data, such as looking at more granular issues within social determinants of health like Native American health or health of children in certain neighborhoods. This application has never been seen before in literature and we hope to use this novel method to find better results than simply finetuning BERT.

To provide a more robust definition of in-context learning, it is essentially a Bayesian framework for “locating” latent concepts acquired by a language model by conditioning the LM simply on an input-output example, rather than optimizing hyperparameters using backpropagation. An input-output example would thus consist of the pairing of a question-answer, such as “Q: Does this person have food insecurity?” and “A: 1” or “A: 0” depending on the true label in the data. In-context learning prompts are a list of concatenated IID (independent and identically distributed) training examples. The diagram below illustrates how this would work as the input to our model:



The main advantage of this approach is that a model can learn using much less data, as the input-output pairs allow for a much faster calibration of the model to respond specifically to the question of whether a person has food insecurity, thus learning prompt associations, rather than the standard next word prediction. Mathematically, this can be viewed as Bayesian inference of a prompt concept shared by every example (in this case, food insecurity). The model is improving the posterior distribution over this prompt concept, as seen by the below notation:

$$p(\text{output}|\text{prompt}) = \int_{\text{concept}} p(\text{output}|\text{concept}, \text{prompt}) p(\text{concept}|\text{prompt}) d(\text{concept})$$

In-context learning is more robust to some noise, particularly if the signal (given as the KL divergence between other learned concepts and the prompt concept) is higher than the noise, allowing the model to distinguish the prompt concept from other concepts quite easily. We conducted hyperparameter tuning to determine the optimal number of test examples to label with input prompts to provide our model, setting the number of examples to 15.

4.3 Results and Analysis

	Precision	Recall	F1 Scores
0 (False)	0.72	1.00	0.84
1 (True)	1.00	0.64	0.38
Accuracy			0.75

Figure 1: Basic model results of classification using in-context learning in BERT model

Since this is a classification task, we decided to use F1 scores and accuracy as our main metrics. The accuracy rate is defined as the (total number of correct predictions / total predictions), and F1 score is the harmonic mean of the precision and recall values, defined as (2 * precision * recall / (precision +

recall)). We wanted to use the F1 score to account for the fact that there are tradeoffs between precision and recall, and using the F1 score appropriately finds the middle ground.

Our results show positive indications that in-context learning performs better than the baseline BERT model (without in-context learning). This new model's accuracy performs better by 10% over the baseline model, where we used only BERT to classify the patients based on the social determinants of health. Our F1 scores for the false values and the true values also increase, from 0.77 to 0.84 for the false values, and from 0.31 to 0.38 for the true values. We reason that this means that our in-context learning model has learned the prompt, or has a general sense of the purpose of this classification. This makes the model more clear on what question we want it to answer. This technique is especially time-efficient and power-efficient because it does not require pretraining the entire BERT model, and instead uses this emergent property of large language models (LLMs) that uses just a few examples to learn the prompt.

One method we could have tried was to improve the recall value on true clinical notes. This tells us that of the full set of patients with food insecurity, we identified 64% of them during our classification task. This means that the model wasn't able to find all of the positive clinical notes we wanted. On the other hand, the recall was high for the false value, meaning that we detected all of the true negatives, meaning the model was good at telling if a patient did not have food insecurity. In the context of our application, we don't need to achieve 100% recall accuracy, since our model is another tool that would aid doctors in identifying food insecure patients, and does not have real time urgency in classifying. We can modify or change the recall values by changing the binary classification threshold on the last layer. On the other hand, our precision values were relatively good.

In analyzing our errors, we did a deep dive on the clinical notes that were incorrectly classified. We divided up the errors based on potential hypotheses for incorrect classification. For example, mentions of the word food but not associations with food insecurity specifically caused the model to classify incorrectly, showing that our association with the prompt specific question needs to be improved. In this specific case, approximately 71% of errors were attributed to this issue. From this, we then thought of how to improve our in-context learning prompting so that the in-context learning layer would understand the question better.

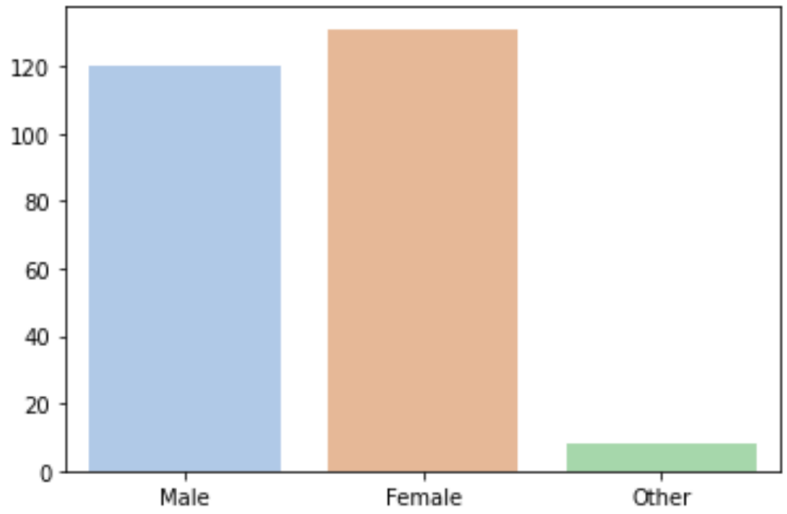
Most importantly, this shows the potential for in-context learning to be used on BERT, a method that has not been seen in the literature before. Previously, in-context learning has been used on GPT-3, a generative model that is autoregressive, and not bidirectional like BERT is. This method being applied on different architectures also shows that this is a general emergent property of all LLMs, and not GPT-3 like architectures. The approach of using GPT-3 instead of BERT may have resulted in different accuracy percentages for a few reasons. GPT-3 doesn't need to be finetuned (there are too many parameters), but we had the choice to finetune BERT. For future directions, we can compare in-context learning performance in GPT-3 vs BERT.

In the field of healthcare, reducing inefficiencies and allowing doctors to identify vulnerable patients will allow for greater patient care and decrease the burden on the doctor. This allows us to best direct social sources towards those who need it most. This model can be applied on all clinical notes of all patients, especially the most vulnerable patients in low-income healthcare settings.

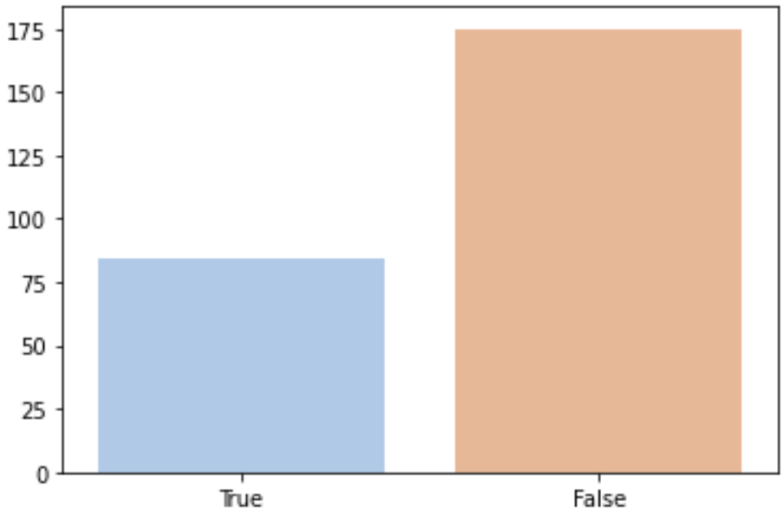
APPENDIX

Dataset Visualizations

Gender Breakdown



Outcome Breakdown



Team Member Contributions

Stanley Yang: Stanley worked on researching various models to use, including LSTM models, Word2Vec, and BERT, understanding the drawbacks and benefits to each. He also worked on constructing the code for our baseline model, as well as the code for our improved model with the in-context learning implementation.

Viraj Mehta: Viraj worked on researching in-context learning and other potential novelties to introduce to the project. He also worked on understanding masking and tokenization for natural language processing. In addition, he contributed to constructing the code for our baseline model, as well as the code for our improved model with the in-context learning implementation.

Jeremy Tian: Jeremy worked on optimizing hyperparameters such as learning rate, number of epochs, and maximum sentence length for tokenization, as well as worked on constructing code for our baseline model, as well as the code for our improved model with the in-context learning implementation. He also contributed to researching BERT models, including Clinical BERT and BioBERT.

Code Link:
https://github.com/viraj28m/CS230-Final-Project/blob/main/CS%20230%20Project%20Milestone%20(Viraj%2C%20Jeremy%2C%20Stanley).ipynb

References

Special thanks to Akshay Swaminathan in the Stanford School of Medicine for his guidance and mentorship, as well as his provision of the labeled STARR dataset.

"Bert 101 - State of the Art NLP Model Explained." *BERT 101 - State of the Art NLP Model Explained*, https://huggingface.co/blog/bert-101#2-how-does-bert-work.

Khurshid, Shaan, et al. "Cohort Design and Natural Language Processing to Reduce Bias in Electronic Health Records Research." *Nature News, Nature Publishing Group*, 8 Apr. 2022, https://www.nature.com/articles/s41746-022-00590-0#Sec8.

Lin, J., Jiao, T., Biskupiak, J. E., & McAdam-Marx, C. (2013). Application of electronic medical record data for health outcomes research: a review of recent literature. *Expert review of pharmacoeconomics & outcomes research*, 13(2), 191–200. https://doi.org/10.1586/erp.13.7.

"Natural language processing in healthcare medical records." *ForeSee Medical. (n.d.)*. Retrieved October 13, 2022, https://www.foreseemed.com/natural-language-processing-in-healthcare.

Patra, Braja et al. "Extracting Social Determinants of Health from Electronic Health Records Using Natural Language Processing: A Systematic Review." *Journal of the American Medical Informatics Association: JAMIA, U.S. National Library of Medicine*, https://pubmed.ncbi.nlm.nih.gov/34613399/.

Ruan, Xiaowen et al. (2021, March 2). "Health-adjusted life expectancy (Hale) in Chongqing, China, 2017: An Artificial Intelligence and big data method estimating the burden of disease at City Level". *The Lancet Regional Health - Western Pacific*. 2 March 2021, https://www.sciencedirect.com/science/article/pii/S2666606521000195?pes=vor.

"Tapping into EHR text fields with NLP." *Prometheus Research Data Management Solutions*. 11 July 2022, https://www.prometheusresearch.com/using-natural-language-processing-nlp-to-make-us-of-ehr-text-fields-for-medical-research/#:~:text=Previous%20Next-,Using%20Natural%20Language%20Processing%20(NLP)%20to%20Make%20Use%20of%20Electronic,data%20on%20a%20massive%20scale.

Xie, Sang Michael and Sewon Min. "How Does In-Context Learning Work? A Framework for Understanding the Differences from Traditional Supervised Learning." SAIL Blog, 1 Aug. 2022, https://ai.stanford.edu/blog/understanding-incontext/.

Xie, Sang Michael et al. "An Explanation of In-Context Learning as Implicit Bayesian Inference." Cornell University, 21 July 2022, https://arxiv.org/abs/2111.02080.

Ye, Xi and Greg Durrett. “The Unreliability of Explanations in Few-shot Prompting for Textual Reasoning.” *36th Conference on Neural Information Processing Systems*, 2022, https://arxiv.org/pdf/2205.03401.pdf.