Nicholas Labuda, Rahul Muthuswamy
April 17th, 2024

Clinical Cohort Extraction Report

This project's goal was to automate the selection of patients for clinical trials through various Natural Language Processing (NLP) methods applied to clinical notes. Ideally, it should be able to label clinical notes as meeting or not meeting certain criteria (also called labels) so that these can be checked against for a number of clinical trials. This is a difficult task because medical language is distinctly different from everyday language, so normal NLP methods often fall short of success. Furthermore, each doctor writes their Electronic Health Records (EHR, synonymous with clinical notes) differently. There is no one standard for format, phrasing, or included information. This complicates the task even further, as the lines for what is correct for a task like determining if a patient meets a criterion get blurry, even for humans. Despite this difficulty, there have been effective methods produced. However, these are often rules-based classifiers and not generalizable to any other task, even ones with the same abstract goal. When n2c2 ran the clinical cohort extraction task in 2018[1], there were some high performing models, with the best achieving a micro-averaged f1 of 0.91[2]. This is certainly a useful model; however, it is not generalizable to any additional or adjusted criterion because it was rules-based. Since 2018, there have been multiple state-of-the-art NLP models released for public use. Thus, we wanted to evaluate some of these models on the task that so far has only seen success with rules-based approaches. In spirit of this, we used the data n2c2 used for their task in 2018.

The clinical notes themselves are split by patient. A patient can have anywhere between three to five notes, each of which were written by doctors and anonymized. These notes were given to annotators, who marked each patient as meeting or not meeting 13 different criteria. The criteria are as follows[3]:

1. DRUG-ABUSE: Drug abuse, current or past
2. ALCOHOL-ABUSE: Current alcohol use over weekly recommended limits
3. ENGLISH: Patient must speak English
4. MAKES-DECISIONS: Patient must make their own medical decisions
5. ABDOMINAL: History of intra-abdominal surgery, small or large intestine resection, or small bowel obstruction.
6. MAJOR-DIABETES: Major diabetes-related complication. For the purposes of this annotation, we define "major complication" (as opposed to "minor complication") as any of the following that are a result of (or strongly correlated with) uncontrolled diabetes: a. Amputation b. Kidney damage c. Skin conditions d. Retinopathy e. nephropathy f. neuropathy
7. ADVANCED-CAD: Advanced cardiovascular disease (CAD). For the purposes of this annotation, we define "advanced" as having 2 or more of the following: a. Taking 2 or more medications to treat CAD b. History of myocardial infarction (MI) c. Currently experiencing angina d. Ischemia, past or present
8. MI-6MOS: MI in the past 6 months

9. KETO-1YR: Diagnosis of ketoacidosis in the past year
10. DIETSUPP-2MOS: Taken a dietary supplement (excluding vitamin D) in the past 2 months
11. ASP-FOR-MI: Use of aspirin to prevent MI
12. HBA1C: Any hemoglobin A1c (HbA1c) value between 6.5% and 9.5%
13. CREATININE: Serum creatinine > upper limit of normal

There are 202 patients in the training data set and 86 in the test data set. Many of the labels are incredibly imbalanced. For example, there is only one patient that meets KETO-1YR in the training dataset, and zero in the test dataset. In order to account for this imbalance during evaluation, a variation of f1 is used that n2c2 called micro-averaged f1. As they put it[4], "For each criterion, we calculated the correct P, R, and F1 for both "met" and "not met" answers, then averaged those to get the micro score for each criterion. Then we averaged all of those together to get the overall micro-averaged F1 score."

The models we created are a logistic regression model as a baseline, a ClinicalBERT[5] model fine-tuned as a multi-label classifier, the same ClinicalBERT model fine-tuned individually for binary classification for each label, zero-shot prompting with ChatGPT, and finally ChatGPT summarization fed into a ClinicalBERT model fine-tuned as a multi-label classifier. Most of the models used required shortening each patient's notes to 512 tokens (median number of tokens per patient was ~3000). For the logistic regression, word embeddings from the ClinicalBERT model were used, as this would provide a good baseline of whether or not the model has additional benefit over a simple logistic regression. It was not evaluated whether these embeddings provided a significant advantage over Word2Vec or a medical Word2Vec. The regressions were trained one for each label, as logistic regression is a binary classification model. They were trained for 20 epochs; after this, there was a dropoff in the amount of loss reduced. Next, The multi-label ClinicalBERT was made by initializing the HuggingFace model as a sequence classifier. This initializes a classifier layer with 9997 params on top of the 108310272 pre-existing, frozen params. This was trained for five epochs as after that there was no significant improvement. We also created individual, binary classification models. We made these by extracting the [CLS] token from the last hidden state, then using this as the input to a neural network. The layers and units of this neural network were played around with a bit, but we settled on 512 units and two layers as this was the least amount of training required without significant dropoff. The networks were trained for five epochs because after this there was no significant improvement. For the zero-shot model, prompting was conducted twice. The first time was done by simply prompting the API with whether or not the note "has" the label. It was then reconducted (with no significant improvement) by prompting the API with the full description of the label's requirements. Finally, the last model prompted the API to summarize the note to the information relating to the requirements, without making conclusions. These summaries were then used to train and evaluate a multi-label ClinicalBERT model in the same method as the previously described multi-label ClinicalBERT.

The results were incredibly underwhelming. A table is included below, but the results will be summarized here. Logistic regression reached a micro-averaged f1 that was usually between 0.51 and 0.54. This is not great, but it is expected for a baseline measure. However, the other models did not do any better. The multi-label classifier was entirely ineffective - it greatly enjoyed picking all 0s or all 1s for each label to minimize loss. For multi-label classification, there is no way I could find to weight the classes individually, so the imbalance data would've had to be under- or over-sampled to account for the balance. However, regardless of this, it simply did not learn from the train data or improve on the test data. It was about as good after two epochs as it was after twenty, only achieving a micro- averaged f1 that hovered around 0.32. This idea stayed true for the individual ClinicalBERT models as well. In the notebook, the last run gave 0.21; such a low number was never seen in my testing but is clearly possible. This number was used in the table as it is the lowest and most recent. In previous testing, they achieved a better micro-averaged f1 that hovered around 0.50. Regardless, the model was unable to learn anything about the training data that generalized to the testing data, as the val_accuracy and val_loss never really changed. In terms of the zero-shot prompting, the API model was able to obtain a slightly higher micro-averaged f1 of 0.64; however, this is still not usable, especially compared to rules-based models. Due to the low max tokens of ClinicalBERT (that being 512), we had the thought to have ChatGPT summarize the note then have ClinicalBERT evaluate it. However, instead of achieving this effect, we combined the worst of both worlds. The ClinicalBERT model trained on these summaries and evaluating summarized notes achieved a micro-averaged f1 of just 0.35. Overall, these results were surprising. We imagined that the state-of-the-art models would fare better on this task. This could be simply due to just how difficult the task truly is; as stated before, medical data is hard to apply NLP methods on. However, there is one avenue we did not get to explore.

	AB	CA	AA	AM	CR	DS	DA	EN	HB	KE	MJ	MK	MI	F1
LR	.58	.56	.49	.54	.57	.48	.49	.43	.51	.50	.61	.48	.47	.52
ML	.39	.26	.49	.40	.14	.36	.49	.42	.20	.50	.33	.48	.48	.37
IM	.17	.26	.02	.40	.14	.26	.02	.42	.20	.00	.25	.48	.05	.21
ZS	.59	.77	.65	.84	.68	.59	.71	.91	.68	.50	.58	.32	.53	.64
SB	.17	.26	.49	.40	.14	.26	.49	.42	.20	.50	.33	.48	.48	.36

LR - logistic regression | ML - multi-label ClinicalBERT | IM - individual model ClinicalBERT | ZS - zero-shot prompting | SB - summarize with ChatGPT, decide with multi-label ClinicalBERT

Within the last year, there was a Longformer[6] published to the public that was pre-trained on clinical notes and medical data. The advantage to a Longformer is its much larger max token count. For example, the clinical one we looked at had a max token count of 4096. After removing stopwords from the clinical notes, this would have been long enough to only truncate a

few of them. In the future, with more time and compute, it would be great to explore the power of this Longformer and see if it fares any better on this task. If it does, then maybe the state-of-the-art is not so bad, and we just chose models ill-equipped to handle a task as difficult as this one. In addition to the clinical Longformer, it would be interesting to explore multishot prompting with ChatGPT. By giving it more information, it might be possible to have it understand the requirements more, significantly improving its results.

Overall, the models we chose to create did poorly on the task. Despite using state-of-the-art models like ClinicalBERT and ChatGPT, they did not outperform a basic logistic regression when deciding whether patients met certain medical criteria. There is still hope, though, as there are some methods we did not try that could achieve better scores.

Works Cited

1. https://n2c2.dbmi.hms.harvard.edu/2018-challenge, challenge overview
2. https://academic.oup.com/jamia/article/26/11/1163/5575392, top performer f1 scores
3. https://huggingface.co/datasets/bigbio/n2c2_2018_track1, label descriptions
4. https://academic.oup.com/jamia/article/26/11/1163/5575392, micro-averaged f1 score
5. https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT, ClinicalBERT model
6. https://huggingface.co/yikuan8/Clinical-Longformer, clinical longformer