Annotating Sparse Risk Factors in Clinical Records with BERT

Stanford CS224N Custom Project

May Jiang
Department of Computer Science
Stanford University
mayjiang@stanford.edu

Abstract

Though there is an abundance of medical information collected in patient clinical records, these records are typically in the form of fragmented free text, such that the task of extracting the relevant pieces can be costly. In this project, we revisit the 2014 i2b2 challenge for identifying risk factors for heart disease in clinical records, focusing on annotating the smoking status and family history of cardiovascular disease, two of the most difficult risk factors in the challenge due to the sparsity of their less common classes. The teams participating in the 2014 challenge applied a combination of hand-written rules and classifiers such as SVM; the objective of this paper is to adapt more recently developed transformer models for this task in order to evaluate the suitability of these models and to understand whether these models can be trained as a substitute for more explicit reasoning in rule-based systems. Fine-tuning BERT, as well as Clinical BERT and BlueBERT – two BERT-initialized models further pre-trained for the clinical and biomedical domains, we find that Clinical BERT and BlueBERT achieve slightly higher F1 scores than BERT, but within margin of error. Moreover, we find that basic oversampling and class weighting approaches to address the class imbalance do not improve the overall performance of the BERT models on this task, as the tradeoff weakens the model's performance on more common classes. The extraction of the span of text within a clinical record most relevant to the risk factor, and the length of the span that is extracted, however, do significantly impact the performance – and for the smoking risk factor, with simple heuristics for extracting the relevant part of a clinical record, BERT models achieve performance comparable to many of the highest scoring models from the 2014 challenge.

1 Key Information to include

- Mentor: Professor Olivier Gevaert (olivier.gevaert@stanford.edu)
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Extracting useful medical information from patient clinical records, which are often only available as fragmented free-form text, is an important but difficult and manually labor-intensive task. The 2014 i2b2 challenge exposed the potential for automatic extraction of risk factors for cardivascular disease from clinical records of diabetes patients, with the highest-performing participating teams achieving F1 scores close to 0.9 on the test data, using a combination of hand-written rules with traditional machine learning classification models such as SVM. In this project, we investigate the performance of recent transformer models - BERT, Clinical BERT, and BlueBERT - applied to the task of identifying the smoking status and family history of cardiovascular disease risk factors in clinical records, through the lens of the 2014 Risk Factor identification challenge on the i2b2 dataset. Further, we consider approaches to address the challenges of data sparsity and class imbalance in the data and how they affect the performance of these models.

Most existing approaches for this task relied heavily on explicit rule-based systems, and though these systems were able to attain strong performance, there were some areas on which the systems had a little more difficulty. In particular, on the 2014 challenge dataset, though the top participating teams achieved F1 scores over 0.9, performance on the risk factor categories that appeared more sparsely in the data was weaker across almost all teams [1].

Since then, the newest deep learning advances in text processing - transformers in particular - have not been thoroughly explored for this task. Since these recent transformer models such as BERT, Clinical BERT, and BlueBERT have been shown to achieve strong performance on similar biomedical NLP tasks, the objective of this paper is to determine whether fine-tuning these models on this task may lead to strong performance as well, and to better understand the strengths and weaknesses of these transformer model variants for the task as compared to more traditional models and hand-written rule-based systems - at a higher level, whether or to what extent these learned neural models can act as a substitute for more human-involved explicit reasoning in the clinical domain.

3 Related Work

The task of annotating clinical records is not a new one; however, prior to the i2b2 challenge there were few shared tasks using clinical texts for training and testing due to the barriers in obtaining and sharing medical data [2]. Other tasks have involved annotating information such as obesity [3] and medications [4]. In these tasks as well as the 2014 challenge, the best performing teams useda combination of classifiers such as SVM and logistic regression with hand-written rules such as regular expression matching for specific phrases [5], and extraction of features such as document structure and medical entities [6]. Only in most recently developed shared tasks have teams begun to take advantage of the newest NLP advances in deep learning models such as BERT [7].

The introduction of BERT [8] was a transformative step in natural language processing. Through pretraining bidirectional representations using a novel ”masked language model” objective and next sentence prediction, the BERT model, pretrained on large general text corpora and fine-tuned for specific tasks, advanced the state-of-the-art in numerous NLP tasks. One of the most notable contributions of BERT was that whereas previous language models were either unidirectional or treated left and right context independently, the ”masked language model” objective enabled BERT to fuse the left and right contexts and ultimately produce a more robust model. Moreover, BERT was one of the first general language models that was able to achieve high performance on a wide variety of tasks simply by fine-tuning. However, because BERT was pretrained on general text - BooksCorpus and English Wikipedia - it was not as well-adapted to different domains that drew from heavily domain-specific vocabularies. The biomedical domain is one such area that uses a large amount of specific technical terminology not typically found in general-purpose text corpora.

BioBERT [9] was the first of its kind to address this problem for the biomedical domain. The authors started with the pretrained base BERT model and pretrained additionally with a large amount of PubMed abstracts and PubMed Central full-text articles. The authors then fine-tuned their BioBERT models for three common biomedical NLP tasks - named entity recognition, relation extraction, and question-answering, and showed a significant improvement over the performance of the general pretrained BERT on these tasks as a baseline.

The authors of Clinical BERT [10] chose to build on these existing models because clinical text, such as physician notes, can differ substantially in vocabulary and linguistic structure compared to both general-purpose text as the books and Wikipedia articles used to pretrain BERT as well as non-clinical biomedical text as the biomedical research abstracts and texts used to pretrain BioBERT. Because there are many existing and potential important applications of natural language processing that involve clinical text, the authors sought to build a model that would improve upon the performance of BERT and BioBERT for these clinical domain applications and to make this model publicly available.

Though Clinical BERT outperforms BERT and BioBERT in some biomedical tasks such as the MedNLI - a natural language inference task, as well as a concept extraction and entity extraction task, in other tasks such as de-identification, it does more poorly. This may be attributed to differences in vocabulary and parameters, or that the authors fine-tune using only clinical data from the intensive healthcare unit of a single healthcare institution [10]. To facilitate model development for a wider variety of biomedical tasks, the authors of the BlueBERT model introduce a benchmark called Biomedical Language Understanding Evaluation (BLUE), with five tasks with varying textual datasets, data sizes, and difficulties [11]. The models they develop are pretrained on a combination of PubMed abstracts and clinical notes, and are shown to outperform a set of baselines including BioBERT and ELMo [11]. In this project, we aim to attain a better understanding of the usability of these models for clinical domain NLP tasks by comparing the performance of the general BERT model with that of the BERT models pretrained on clinical texts - Clinical BERT and BlueBERT – on the specific task of extracting sparse risk factors from clinical records.

4 Approach

In this project, we fine-tune BERT, Clinical BERT, and BlueBERT to identify two sparse risk factors for cardiovascular disease, smoking status and family history of cardiovascular disease. The task is challenging not only because of the relatively small amount of data – a common problem with clinical data in practice due to privacy issues and the costliness of manual annotation – and the complex nuances of the annotation rules, but also because the class distributions of the labels are significantly skewed, as discussed in the Data section. Further, the clinical records are often long and noisy – the parts relevant to each risk factor vary in length but typically comprise only a relatively small span within a clinical record, and the phrases, abbreviations, syntax, sections, and overall document structure vary across documents.

BERT, Clinical BERT, and BlueBERT are pretrained embedding models. For classification the final hidden state **h** of the first token is taken as the representation of the whole sequence, and then a simple softmax classifier is used to predict the probability of the label c as

$$p(c|\mathbf{h}) = softmax(W\mathbf{h})$$

and W and BERT parameters are fine-tuned to maximize the log-probability of the correct label [12]. This is done with BertForSequenceClassification and AutoModelForSequenceClassification [13].

One drawback of directly applying a BERT model to the clinical records is that the length of the clinical records often significantly exceeds the maximum length that BERT can handle. The maximum length of an input sequence to BERT is 512 tokens; however, the clinical record texts are often about 1000 tokens, so that applying a BERT model directly results in the clinical record text being truncated to only the first 512 characters. Since the description of a patient's smoking status often does not occur at the beginning of the clinical record, this leads to a loss of critical information for determining the smoking status.

To overcome this, we implement heuristic extraction functions to extract a continuous span of n characters from the clinical record most likely to be relevant to the risk factor based on a priority list of keywords. These keywords were determined based on examining documents in the training data set. For smoking status, the extraction function priority-matches on the keywords "smok", "tobacco", "cigar", "tob", "sh", and "social history", and for family history of CAD, the extraction function priority-matches on the keywords "family history", "family", "fhx", "mother", "father", "sister", "brother", "fh", "sh", and "social history". If none of the keywords is found in the text, a random n-character span of the text is chosen.

In addition, we consider basic approaches to address the problem of class imbalance, using the commonly used techniques of specifying class weights and oversampling less frequent classes [14]. Though BertForSequenceClassification and AutoModelForSequenceClassification do not accommodate class weights, we implement class weights by modifying the training process to use CrossEntropyLoss with specified weights rather than the internal loss implementation of those models. Though there are several ways to formulate class weights that set higher weights for less frequent classes and lower weights for frequent classes, on this task they made little difference, so the computations used to produce the models that are reported in the Results section are

$$w(c_x) = 1 - len(c_x)/len(total)$$

where $w(c_x)$ is the class weight for class x, $len(c_x)$ is the number of records in class x in the set and $len(total)$ is the total number of records in the set. The oversampling is done by selecting for each class a random sample, with replacement, of length equal to the number of data records in the largest class.

5 Experiments

5.1 Data

The dataset used in the 2014 i2b2 challenge is split into training, validation, and test sets, divided into parts of approximately 40%-20%-40%. The corpus consisted of 1,304 records representing 296 patients. The records are in the form of unstructured text in XML files with the gold label tags. From these files, the smoking status and family history labels is extracted from the tags, and files missing the label of interest are discarded. Table 1 and 2 display the distribution of each label in the training, validation, and test sets, for smoking status and family history, respectively.

The smoking status annotation consists of five mutually exclusive labels – Current, Past, Ever, Never, and Unknown. A patient's clinical record is annotated as Current for the smoking status risk factor if the patient is currently a smoker or quit less than a year ago – for instance, as indicated by the phrases "Patient says trying to quit 1pack/day habit" or "quit 6mos ago". Meanwhile, the Past label is used when a patient used to smoke but quit over a year ago, the Never label is used when a patient has never smoked, and the Ever label is used when it is unclear whether the patient has quit or there is no information on the timeframe when the patient quit smoking – for instance, if the description is "quit smoking" or "remote history of tobacco dependence". If there is no information about smoking status, the clinical record is labeled as Unknown; the class proportions for smoking status are heavily skewed and this is the most frequent class in the data by a significant margin [1].

The family history annotation is an indicator that is 'present' if the patient has a first-degree relative (parents, siblings, or children) who was diagnosed prematurely (younger than 55 for male relatives, younger than 65 for female relatives) with CAD, and 'not present' otherwise. For instance, clinical records with phrases such as "Father diagnosed w CAD at 49" or "Fam.hist. significant for premature CAD" would be annotated as 'present' while phrases such as "Father w/ CAD, died at 82 yrs", "No known relatives with CAD", or "Both grandfathers prem. CAD", should not be annotated for the risk factor. The vast majority of the clinical records are annotated as 'not present' [1].

Data	Count	Unknown (%)	Never (%)	Past (%)	Current (%)	Ever (%)
Training set	511	0.47	0.25	0.20	0.07	0.01
Validation set	260	0.50	0.22	0.18	0.09	0.01
Test set	511	0.47	0.23	0.22	0.06	0.01

Table 1: Distribution of Smoking Status Labels

Data	Count	Not Present (%)	Present (%)
Training set	521	0.98	0.02
Validation set	269	0.95	0.05
Test set	514	0.96	0.04

Table 2: Distribution of Family History of CAD Labels

5.2 Evaluation method

The micro-averaged F1 score is used to evaluate the models, to be consistent with the evaluation metric from the 2014 i2b2 challenge. Note that F1 is the same as the precision and recall when micro-averaging in a multi-class problem. Accuracy is considered as well, for each of the five classes for the smoking factor, or each of the two classes in the case of the family history factor. The micro-F1 scores of the models trained in this work are compared against those of the top performing teams in the challenge, as well as a baseline of predicting the most frequent class in the training data – 'unknown' for smoking status and 'not present' for family history of cardiovascular disease.

5.3 Experimental details

The code for loading the data, building the models, and training and testing utilize the Huggingface transformers library [13] and adapt code from [15]. The BERT model used is the base, uncased model [8], the Clinical BERT model is the model initialized from BioBERT and trained on all MIMIC notes [10], and the BlueBERT model is the uncased model pretrained on PubMed and MIMIC clinical notes [11]. Each of the models was trained for 10 epochs, and the final model is selected selected from the epoch with the lowest validation set loss. A learning rate of 1e-5 was used with the Adam optimizer [16], and a batch size of 3 is used. The training took about an hour for each model in the Azure environment.

5.4 Results

The results of the trained models on the test data are displayed in tables 3-5 for the smoking status risk factor, including median micro-F1 scores with their 95% confidence intervals for all model variants and micro-F1 scores from the baseline and from top teams in the 2014 challenge as reported. Table 6 reports the accuracies by class of smoking status for a subset of these models. The BERT variant used is indicated in the tables, along with the number of characters extracted; for instance, BERT-800 denotes a model initialized with the original BERT base model, and fine-tuned and tested on 800-character extracts from the clinical records, extracted as described in the Approach section.

Each of the model approaches varying the extraction length, weights and oversampling, and using BERT, Clinical BERT, and BlueBERT, was also trained and tested for the family history risk factor. However, in all but a few cases, the models produce identical F1 scores to that of the baseline that predicts the most frequent class – family history not present – in all cases. Thus, Table 7 displays the micro-F1 scores and class accuracies where available for only a selected set of these models.

We find that for both risk factors, the highest scoring model from the 2014 challenge was still the result with the highest F1 score, but that the best performing BERT models achieve F1 scores on par with several of the top teams in the challenge. As shown in Tables 5 and 7, the best performing BERT, Clinical BERT, and BlueBERT models all achieve F1 scores that would rank within the top 5 highest scores from the challenge for both smoking status and family history of CAD. The extraction function makes a significant difference for performance on annotating smoking status, improving the models from performance comparable to the baseline of always predicting the most frequent class – a micro-F1 score of close to 0.475 – to highest scores of 0.865 for BERT with extracts of 1000 characters, 0.881 for Clinical BERT with extracts of 600 characters, and 0.867 for BlueBERT with extracts of 600 characters. Class weights and oversampling, however, in most cases do not appear to make any difference for the performance of the models outside of the margin of error. On average, the scores of the models with class weights and oversampling tend to be lower than those of the models without any adjustment, with the exception of BlueBERT.

Number of characters extracted	BERT	Clinical BERT	BlueBERT
No extraction	0.473 (0.432, 0.518)	0.475 (0.434, 0.520)	0.476 (0.434, 0.520)
400	0.828 (0.795, 0.861)	0.846 (0.811, 0.879)	0.848 (0.814, 0.879)
600	0.863 (0.832, 0.893)	**0.881 (0.850, 0.906)**	**0.867 (0.834, 0.895)**
800	0.865 (0.836, 0.895)	0.865 (0.832, 0.873)	0.830 (0.797, 0.863)
1000	**0.865 (0.836, 0.900)**	0.846 (0.832, 0.895)	0.844 (0.811, 0.873)
1200	0.857 (0.824, 0.887)	0.834 (0.803, 0.865)	0.816 (0.783, 0.848)
1400	0.820 (0.787, 0.854)	0.826 (0.791, 0.859)	0.806 (0.773, 0.842)

Table 3: Smoking classification micro-F1 scores with 95% confidence interval, best result for each BERT model bolded

6 Analysis

The reason that the models achieve stronger performance when the extraction is used is intuitive: because BERT simply uses the first 512 tokens when in many cases the smoking status is described

	No Adjustment	Class Weights	Oversampling
BERT-800	**0.865 (0.836, 0.895)**	0.863 (0.832, 0.893)	0.840 (0.807, 0.869)
BERT-1000	**0.865 (0.836, 0.900)**	0.863 (0.834, 0.893)	0.828 (0.795, 0.863)
Clinical BERT-600	**0.881 (0.850, 0.906)**	0.875 (0.844, 0.902)	0.877 (0.846, 0.904)
Clinical BERT-800	**0.865 (0.832, 0.873)**	0.848 (0.814, 0.879)	0.836 (0.803, 0.805)
BlueBERT-400	0.848 (0.814, 0.879)	0.842 (0.807, 0.871)	**0.859 (0.828, 0.889)**
BlueBERT-600	0.867 (0.834, 0.895)	**0.873 (0.844, 0.902)**	0.863 (0.834, 0.893)

Table 4: Smoking classification micro-F1 scores with 95% confidence interval for best performing models, best result across class imbalance approaches for each model bolded

	micro-F1
BERT-1000	0.865 (0.836, 0.900)
Clinical BERT-600	0.881 (0.850, 0.906)
BlueBERT-600-weighted	0.873 (0.844, 0.902)
2014 challenge, 1st	**0.916**
2014 challenge, 5th	0.854
2014 challenge, 10th	0.815
Most frequent class (‘unknown’)	0.475

Table 5: F1 scores of highest scoring BERT, Clinical BERT, and BlueBERT models and baselines, best result bolded

	Current	Ever	Never	Past	Unknown
Clinical BERT-no extraction	0	0	0	0	1
Clinical BERT-600	0.183	0	0.926	0.859	0.971
Clinical BERT-600-weighted	0.303	0	0.899	0.742	0.955
Clinical BERT-600-oversampled	0.574	0	0.926	0.776	0.950
Clinical BERT-1000	0.210	0	0.841	0.807	0.959
BERT-1000	0.480	0	0.877	0.753	0.976
BlueBERT-600	0	0	0.934	0.858	0.966
BlueBERT-600-weighted	0.092	0	0.934	0.840	0.975
BlueBERT-400	0	0	0.934	0.735	0.983
BlueBERT-400-oversampled	0.424	0	0.866	0.753	0.975

Table 6: Accuracies by class of smoking status, reported for selected models

toward the middle of the clinical record and the order of the sections within a clinical record varies significantly across records, the part of the clinical record relevant to the classification is often lost to the model that does not use extraction. Inspecting one straightforward clinical record for which the BERT, Clinical BERT, and BlueBERT models using the extraction function all correctly determine that the patient was a past smoker from the phrase “The patient quit smoking 20 years ago but had smoked 1.5ppd*20yrs (30 pack year)”, the corresponding models without extraction fail to correctly classify this example simply because the phrase does not occur until almost the 600th token of the clinical record.

The reason that the models vary in performance depending on the length of the span extracted, however, is less clear. Table 3 shows the performance of the BERT variant models with the extraction function applied to the data, with extracts of different lengths in number of characters. Each of the three models appears to benefit from increasing the number of characters extracted from 400 to 600, but the Clinical BERT and BlueBERT models’ F1 scores decrease slightly when the number of characters extracted is increased from 600 to 800. This suggests that either there may be a significant number of phrases that are longer than 400 characters but less than 600 characters long, or that 600

	micro-F1	Accuracy: Not Present	Accuracy: Present
BERT-no extraction	0.963 (0.947, 0.981)	1	0
BERT-1200	0.963 (0.949, 0.979)	1	0
BERT-600	0.965 (0.947, 0.981)	0.964	0.108
BERT-600-weighted	0.965 (0.949, 0.979)	0.965	0.052
BERT-600-oversampled	0.963 (0.944, 0.979)	1	0
Clinical BERT-600	0.963 (0.947, 0.979)	1	0
BlueBERT-600	0.963 (0.946, 0.979)	1	0
2014 challenge, 1st	**0.981**	-	-
2014 challenge, 5th	0.963	-	-
2014 challenge, 10th	0.949	-	-
Most frequent class (‘not present’)	0.963	1	0

Table 7: Performance metrics for annotating Family History of CAD, selected models and baselines

characters is enough to capture relevant data that may be missed by the basic extraction heuristic, that these models benefit from longer spans longer than 400 but may be weakened when adding characters with additional noise beyond 600 characters. In the case of the base BERT, the models achieve comparable performance for 600, 800, and 1000 characters. In one patient record, one of the first sections, “past medical history", contains a mention of ‘smoker’: “past medical history: cad, history of stemi in 2077, eight stents including lad, at least x 2, biv icd placement , last cath at och showed multivessel disease, biv icd, ddd st. jude, 05/13/2081, chf, diabetes, hypertension, former smoker." However, there would not be enough information to determine that the patient was a Past smoker rather than an Ever smoker until a later sentence a few brief sections after that: “social history: he is a retired purchasing agent, quit smoking a few years ago, had smoked one pack per day. he has used no alcohol or illicit drug use, a very supportive family." As a result, this information is just barely captured in the 600 character extract, and is comfortably contained in the 1000 character extract, but not present in the 400 character extract. The BERT-600, BERT-800, BERT-1000, and BERT-1200 models all correctly classify this patient record as Past, but the BERT-0 and BERT-400 do not. Another way to address this issue in future work would be to pass an extract for each keyword match to the model, rather than only the first time a keyword is found, and to take the maximum or average of the results on the each of the extracts to classify the document.

To mitigate the class imbalance in the data, we used class weights and oversampling approaches. However, these approaches did not improve the overall F1 scores of the models, with the possible exception of BlueBERT, as displayed in Table 4. This can be attributed to the way that the overweighting or oversampling – which are conceptually the same approach, to induce less frequent classes to factor more heavily into the loss function – affects different classes. Looking at Table 6, we see that the weighting and oversampling improves the accuracy of classifying the records as Current in every case – for Clinical BERT-600, only 18.3% of the Current labels are correctly extracted but with weighting and oversampling 30.3% and 57.4% of the labels are correctly identified. Likewise, for BlueBERT-600, weighting increases the accuracy on Current from 0 to 0.092, and for BlueBERT-400, from 0 to 0.424. The difference in overall performance, however, is shaped by the tradeoff of accuracy on Current for less accuracy on more common labels. In the case of Clinical BERT with 600 characters, the shift of emphasis toward correctly classifying Current leads to a significant decline in performance on the more common classes of Past and a smaller but noticeable decline in performance on Unknown, the most common class. For BlueBERT, there is a much smaller decline in performance in the Past and Unknown classes. The reason for this is unclear, but one possibility is that it is due to BlueBERT having more sources of pretraining data or preserving more of the BERT parameters and the potentially better generalizability could make the model more robust to differences in class distributions.

Overall, the highest-scoring Clinical BERT and BlueBERT models outperform the base BERT model, though within margin of error. This is as expected, since Clinical BERT and BlueBERT are pretrained specifically for the biomedical domain, and would have better exposure to clinical syntax and vocabulary. The class accuracies in Table 6 suggest that this may be attributable to BERT performing particularly poorly on Never and Past classes in comparison to the other models. We find that in 11 of the 14 cases that BERT-1000 misclassifies a record that should be labeled as Past, the model misclassifies the record as Ever, and in 15 of the 30 cases that BERT-1000 misclassifies a record that should be labeled as Never, the record is misclassified as Past. It may be that familiarity with clinical vocabulary is advantageous for navigating the nuances of these similar labels.

The performance on the family history risk factor is displayed in Table 7 for a selection of models. Both most of the transformer models and most of the top models in the 2014 challenge achieve only F1 scores equal to that of predicting 'not present' for all cases – these models, even including the Clinical BERT and BlueBERT models with weights and oversampling, are unable to overcome the challenge of the significant class imbalance in the data, a sparsity issue compounded by the large variety of possible descriptions of family history in the clinical records and the complex definition of the family history annotation as described in the Data section. Since family history is less dependent on clinical vocabulary and biomedical terms or syntax, Clinical BERT and BlueBERT do not have an advantage for this risk factor, and the BERT models with 600 characters achieve higher F1 scores by attaining non-zero accuracy on the Present label.

7 Conclusion

In this project, we fine-tuned recent transformer models – BERT, Clinical BERT, and BlueBERT – to the task of identifying smoking status and family history of CAD in clinical records, using the dataset of the 2014 i2b2 challenge. Though the small amount of data, the complexity of the annotation rules, the length of the documents, and the class imbalance and sparsity of uncommon class labels prove challenging for the task, and none of the transformer models outperform the highest scoring models in the 2014 challenge, that these BERT models achieve performance on par with several of the top scoring models that utilize a combination of hand-written rules, feature extraction, and traditional classifiers, is promising.

A major limitation of the models developed in this project is that neural models such as transformers rely on a large amount of data to learn such sparse patterns as the annotation rules for these risk factors, and the most successful system in the 2014 challenge, in addition to external lexicons, hand-written rules, and SVMs, relied on additional annotations as well. Future work could improve upon these models by training using more data – perhaps by annotating the risk factors in and extending to a larger dataset such as MIMIC, a large set of unlabeled clinical data. Other approaches could extend our analysis to improve the extraction heuristics, or to try more complex methods for addressing class imbalance. Ultimately, though transformer models have been little-explored in the biomedical NLP domain to date, given the findings of this work and the strength of BERT models' performance with only a simple extraction function, there is potential for systems and ensembles to leverage BERT for this task and other challenging tasks in the clinical domain as well.

References

[1] Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2. *Journal of biomedical informatics*, 58:S67–S77, 2015.

[2] Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions, 2011.

[3] Özlem Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–570, 2009.

[4] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.

[5] Manabu Torii, Jung-wei Fan, Wei-li Yang, Theodore Lee, Matthew T Wiley, Daniel Zisook, and Yang Huang. De-identification and risk factor detection in medical records. In *Seventh i2b2 Shared Task and Workshop: Challenges in Natural Language Processing for Clinical Data*, 2014.

[6] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24, 2008.

[7] Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[10] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[11] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65, 2019.

[12] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification?, 2020.

[13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[14] Ajinkya More. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*, 2016.

[15] Susan Li. Multi class text classification with deep learning using bert, Aug 2020.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.