Predicting Emergency Department Disposition from Radiology Reports

Stanford CS224N Custom Project

Karen Garcia
Department of Computer Science
Stanford University
karengar@stanford.edu

Yue Andy Zhang
Department of Computer Science
Stanford University
yuezha@stanford.edu

Abstract

In this project, we propose analyzing emergency department disposition through the lens of the radiology reports. Specifically, we have two objectives, conduct an unsupervised multi-label classification to extract medical conditions, followed by supervised classification to predict patients' disposition.

Several unsupervised label extraction methods were used to obtain medical conditions by considering CheXbert labels as a ground truth. We evaluate methods for supervised classification of the disposition by combining the extracted medical conditions and the radiology reports. Our work demonstrates superior performance of using a pretrained BART-MNLI model for zero-shot label extraction. We find that models trained using the extracted medical conditions together with medical annotations outperforms the other methods to predict patients' disposition.

1 Key Information to include

- Mentor: Abhinav Garg
- External Collaborators (if you have any): David Kim, MD PhD, Assistant Professor, Emergency Medicine
- Sharing project: No

2 Introduction

Radiology reports are a key input used by physicians to decide the disposition of emergency department patients. This project seeks to create an automated system to predict emergency department disposition, to assist doctors in timely decision-making.

There are two tasks relevant to disposition prediction from radiology reports:

1. Extract human-readable labels from radiology reports. These labels indicate the presence or absence of medical conditions, eg. pneumonia. There are 39 conditions, and each report may discuss multiple conditions.
2. Predict the patient's disposition: Admit to Inpatient, Discharge, Observation and CDU Observation, based on the text report and/or extracted condition labels.

We propose and compare systems to perform zero-shot/unsupervised label extraction and supervised disposition prediction.

The ED disposition dataset provides ground-truth labels for disposition, but not for the 39 medical conditions. Thus, the first task of label extraction can be framed as unsupervised/zero-shot multi-class classification.

We for zero-shot label extraction, we compare Lbl2vec, a clustering-based method, with pretrained language models. In this case, we use the BART model Lewis et al. (2019) pretrained on MultiNLI dataset Williams et al. (2018).

In this project, we also investigated the effect of different scenarios on predicting the four patients disposition. Specifically, we will address the following questions:

A Model effectiveness using only text report: How well can a model trained on text report perform on patients disposition prediction?

B Model effectiveness augmenting the disposition classifier input with zero-shot labels: Can the extracted condition labels using unsupervised methods together with text reports increase the performance of the model?

C Model effectiveness base on different BERT implementations: Could the dataset on which a BERT model has been pre-trained affect performance?

We compared accuracy, AUC, precision, recall, and F1 score in each scenario. In each case, we use the regular BERT and its biomedical versions.

3 Related Work

3.1 Medical report labeling

Prior methods to automatically extract labels from medical reports fall under two categories. The first category consists of rule-based labelers that use feature engineering built by experts. CheXpert is a rule-based labeler to extract labels from chest radiology reports Irvin et al. (2019). The second category consists of transformers models, which typically do not take advantage of existing feature-engineered labelers.

The CheXbert labeler is a hybrid approach that uses transformer models pretrained on medical corpus, and then fine-tunes the model on the outputs of rule-based labelers and expert annotations, to achieve accurate automated radiology report labeling. The CheXbert labeler performs at accuracy close to human experts. Smit et al. (2020)

Prior work on medical reports labeling typical poses the task as supervised multi-label classification. On the other hand, our dataset is unlabeled (only the disposition label is provided), so we use unsupervised methods for label extraction.

3.2 Unsupervised label extraction

Lbl2vec Lbl2vec is a method for unsupervised document classification. This method first embeds documents and labels in a joint embedding space, and then clusters embeddings using cosine similarity to assign a label to each document (Schopf et al., 2021) (Schopf et al., 2023b). Embeddings are produced using doc2vec or transformer-based language models, where transformer embeddings typically higher quality representations of the input. (Schopf et al., 2023a). The Lbl2vec algorithm assigns a single label for each document, so it cannot be applied out-of-the-box for for multi-label classification.

Zero-shot text classification with transformers Pre-trained models for natural language inference (NLI) can be used as sequence classifiers. To reformulate classification as an NLI task, the text to be classified is the NLI premise, while each candidate label is constructed into an NLI hypothesis, ie. "This report discusses pneumonia". The probability of entailment/neutral/contradiction produced by an NLI model can thus used for classification. Since each candidate label is evaluate independently, this method extends naturally to multi-label classification. (Yin et al., 2019)

3.3 Supervised disposition prediction

Supervised machine learning algorithms have been a dominant method in the data mining field. Patient's disposition prediction using radiology reports is a potential application area for these methods. In this project, we develop an end-to-end process of fine tuning highly robust natural language

processing (NLP) models using Transformers architecture. Bidirectional Encoder Representations from Transformers (BERT) word embedding models have been successfully used for many NLP tasks (Devlin et al., 2019). However, there are many more linguistically complicated concepts in healthcare documentation, often reflecting medical decision-making processes or complex patient characteristics, where performance of transformer-based models has not been as well investigated. Furthermore, the dataset on which a BERT model has been pre-trained could affect performance. One of the objective of this study is to compare performance of regular BERT and its biomedical versions on patience's disposition prediction.

3.4 BERT implementations

In this study were used three BERT implementations. The first one is the base BERT (Devlin et al., 2019). This model uses a vocabulary for English extracted from the Wikipedia and BooksCorpus. Text inputs have been normalized the "cased" way, meaning that the distinction between lower and upper case as well as accent markers have been preserved. The second is base on BioBERT. The original BioBERT was initialized with weights from the base BERT (Lee et al., 2019), and then pretrained on PubMed abstracts and PubMed Central full-text articles. Our model is a fine tuned version of BioBERT on NCBI disease corpus and on the TAC 2017 dataset. The third model was ClinicalBERT (K. Huang, 2019), which is pre-trained on Medical Information Mart for Intensive Care III (MIMIC-III) (A.E. Johnson, 2016).

4 Approach

4.1 Unsupervised label extraction

We compare three methods for unsupervised/zero-shot label extraction, and the labels produced by the best method are used as inputs to the disposition classifier (see purple box in figure 1). The CheXbert labeler is used to evaluate the three methods.

4.1.1 CheXbert

The CheXbert labeler produces 7 of the 39 medical conditions of interest. We run the CheXbert model with the radiology reports as inputs, to obtain labels for these 7 conditions (the other CheXbert labels are discarded). For each condition, the CheXbert output is either positive, negative, uncertain, or blank. We consider labels produced by CheXbert as ground truth, to evaluate and fine-tune the unsupervised classification methods below.

4.1.2 Lbl2vec

The first method used for unsupervised label extraction is a modified version of the lbl2vec algorithm. As input to lbl2vec, we provide the set of 39 medical conditions as labels, and the radiology documents to be labelled. The vanilla lbl2vec algorithm computes cosine similarity scores between each document/label pair, and assigns the most likely label to each document. To support multiple labels for each document, we use the scores directly, and implement a Naive Bayes decision rule for evaluation (see Evaluation section for details).

For disposition prediction, we use either the document/label similarity score, a 39-dimensional vector with values between 0 and 1 indicating the probability of each medical condition. Alternatively, we apply a threshold to each score to get a boolean vector.

4.1.3 Pretrained NLI Model

The second method for unsupervised label extraction is to apply a pretrained NLI model. Each radiology report serves as a premise. For each premise, hypotheses are constructed from each of the 39 labels, like 'This example is <label>'. We use a BART-large model that has already been fine-tuned on the MultiNLI dataset (BART-MNLI), available as bart-large-mnli from Huggingface.

4.1.4 Fine-tuned NLI Model

We also fine-tune the BART-large NLI model using the premise and hypothesis defined previously. We use CheXbert output for the 7 relevant conditions as entailment labels for fine-tuning.

4.2 Disposition prediction methods

In the first method, the goal is to test the effectiveness of the models by training using text reports of the radiology document (see figure 1). The second method or baseline method aims to test a model, using the results of the zero-shot labeler. The third method combine the text reports together with the result of the zero-shot labeler in order to make disposition predictions.

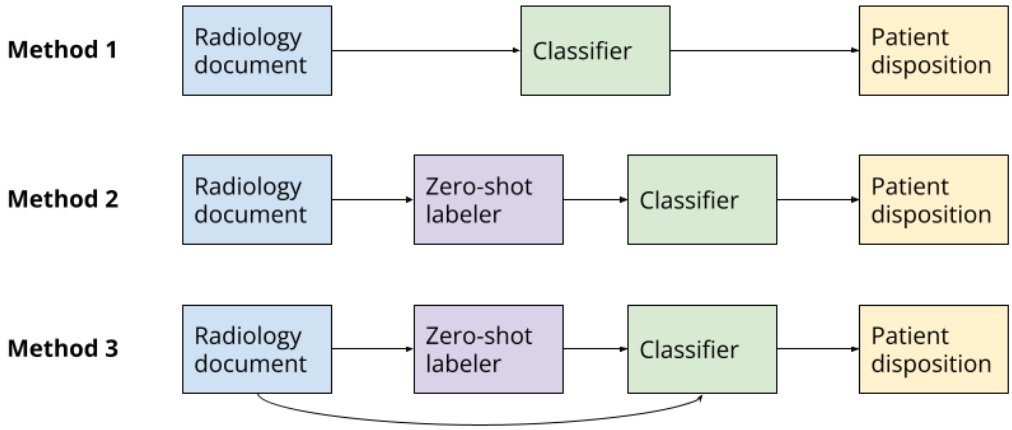

Figure 1: Experiments conducted in the project.

For methods 1 and 3, we fine-tuned several BERT, BioBERT and ClinicalBERT models.

For method 2, we trained multilayer perceptron classifiers, where the input is the 39-dimension score vector for medical conditions. We performing grid search over hyperparameters such as the number of hidden units and learning rate.

5 Experiments

5.1 Data

The dataset consists of 150,000 emergency department records from Stanford Hospital. The key components of each record are:

1. **Report:** a document describing the radiology results, up to a few thousand words
2. **Disposition:** the patient's outcome, determined by a physician. The four categories are Admit to Inpatient, Discharge, Observation and CDU Observation.

Label	Training Dataset (N = 102 304)	Test Dataset (N = 14 615)	Validation Dataset (N = 29 230)
Admit to Inpatient	46 484 (45.44%)	6 608 (45.21%)	13 217 (45.23 %)
Discharge	49 897 (48.77%)	7 187 (49.18%)	14 295 (48.91%)
Observation	5 260 (5.14%)	733 (5.01%)	15 37 (5.26%)
CDU Observation	663 (0.65%)	87 (0.60%)	181 (0.62%)

Table 1: Characteristics of training, test and validation datasets.

Each record also has some additional metadata, such as which radiology procedure performed. We did not use the metadata, so that the results here can be solely attributed the methods' ability to understand report documents. This metadata likely has predictive value and would be used in a real-world system. The number of documents in the training datasets was 102 304, on test data set 14 615, and on validation dataset 29 230 (see Table 1). The dataset has been split by preserving the percentage of samples for each class.

5.2 Evaluation method

For unsupervised label extraction, we evaluate the methods by comparing the predicted labels of each method against the labels produced by CheXbert, which we consider as ground truth. Each method outputs a 39-dimension vector with values between 0 and 1. The elements indicate the independent probability that each medical condition is described as being present in the radiology report. For the seven conditions that are also produced by CheXbert, we learn a decision boundary using Gaussian Naive Bayes, and use this threshold to convert the probabilities into true/false labels. The predicted labels are compared against CheXbert labels for a validation set, to obtain metrics like AUC, accuracy, and F1 score. This method was applied for lbl2vec, the pre-trained BART-MNLI model, and the fine-tuned BART-MNLI model.

For disposition prediction, the dataset contains ground-truth labels for each radiology report. Thus, we can directly compare the validation set output of each classifier with ground-truth labels to metrics like AUC, accuracy, and F1 score.

5.3 Experimental details

Training and evaluation for all models was done using a single Nvidia T4 or A10G GPU on AWS.

5.4 Unsupervised label extraction

To extract labels, the lbl2vec algorithm was run using embeddings from the all-MiniLM-L6-v2 pretrained model, and default hyperparameters. Evaluation took approximately 6 hours. For the NLI methods, we used bart-large-mnli (BART-MNLI). We attempted to fine-tune all the weights and fine-tune just the classification-head. Fine-tuning the classification head was done for 1 epoch, taking approximately 9 hours. Fine-tuning the entire model was done for done for 0.5 epochs and took a similar amount of time.

5.5 Disposition prediction methods

In method 1 and 3, we freeze the BERT and its biomedical versions model and train a randomly initialized linear layer for the classification task. More than 10 cycles of a parameter sampler were conducted for each scenario. ParameterSampler generate parameters from given distributions. Hyperparameters that were optimized for each model and method are listed in the following table:

- Learning rate: log uniform(1×10^{-2},1×10^{-6})
- Dropout probability: 0.0, 0.1, 0.2, 0.3
- Epochs: 3, 5, 10
- Batch size: 16, 32, 64

The model was trained using AdamW optimizer and cross-entropy loss. Each model takes about one to six hours to train, depending on batch size, number of epochs and learning rate.

For the multilayer perceptron classifier used in method 2, grid search was performed over the following hyperparameter space:

- Learning rate: 1×10^{-2}, 1×10^{-3}, 1×10^{-4}, 1×10^{-5}
- Number of hidden layers: 1, 2
- Hidden layer dimension: 16, 64, 256
- Dropout probability: 0, 0.2, 0.4

- Epochs: 10, 100, 500

The model was trained using AdamW optimizer and cross-entropy loss. Hyperparameter search took 5 hours to train all the models.

5.6 Results

5.7 Unsupervised label extraction

We observe that for label extraction, the pretrained BART-MNLI model performs better than lbl2vec for all seven labels (see Table 2). This is most evident when comparing the AUC and F1 score for each label (see Figure 2 for AUC comparison). The accuracy metrics generally quite high because of label imbalance, since it is far more common for a medical condition to be absent from a report, than to be present (see Table 3).

For subsequent experiments, we use the output of the pretrained BART-MNLI model as input to disposition classifiers.

Label	Model	AUC	Accuracy	F1 score	Precision	Recall
Edema	Lbl2vec	0.62	0.96	0.38	0.66	0.26
	BART-MNLI	0.93	0.97	0.74	0.64	0.88
Cardiomegaly	Lbl2vec	0.59	0.98	0.29	0.97	0.17
	BART-MNLI	0.78	0.97	0.45	0.37	0.57
Pneumonia	Lbl2vec	0.59	0.97	0.27	0.58	0.17
	BART-MNLI	0.70	0.94	0.38	0.34	0.43
Atelectasis	Lbl2vec	0.62	0.96	0.32	0.40	0.26
	BART-MNLI	0.68	0.96	0.36	0.34	0.39
Pneumothorax	Lbl2vec	0.53	0.99	0.09	0.20	0.06
	BART-MNLI	0.88	0.97	0.39	0.26	0.79
Pleural Effusion	Lbl2vec	0.72	0.94	0.56	0.77	0.44
	BART-MNLI	0.78	0.94	0.62	0.68	0.58
Fracture	Lbl2vec	0.50	0.89	0.00	0.00	0.00
	BART-MNLI	0.93	0.93	0.76	0.65	0.91

Table 2: Validation metrics for the lbl2vec and BART-MNLI pretrained models applied to label extraction, evaluating against CheXbert output as ground-truth

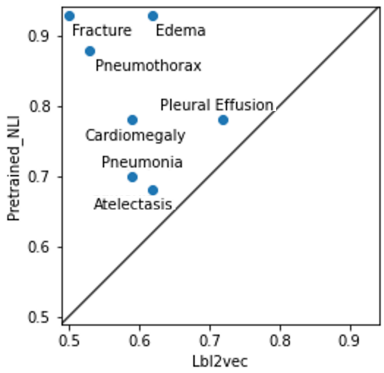

Figure 2: Comparing AUC from Table 2

Label	Present	Absent
Edema	6925 (4.7%)	139224 (95.2%)
Cardiomegaly	3101 (2.1%)	143048 (97.9%)
Pneumonia	4043 (2.8%)	142106 (97.2%)
Atelectasis	4732 (3.2%)	141417 (96.8%)
Pneumothorax	1320 (0.9%)	144829 (99.1%)
Pleural Effusion	11706 (8.0%)	134443 (92.0%)
Fracture	16111 (11.0%)	130038 (89.0%)

Table 3: Distribution of each label, based on CheXbert output, N=146149.

In Figure 3, we compare the prediction distributions for the two label extraction methods. The x-axis is the probability that the condition is present assigned by the model, while the y-axis is the number of reports in that bin. The left column are predictions from the pre-trained BART-MNLI model, and the right column is from lbl2vec. We observe that the pre-trained BART-MNLI model is much more effective at discriminating between true positives and true negatives, with very different distributions. On the other hand, the predicted distribution from lbl2vec are mostly overlapping, resulting in weak classification performance.

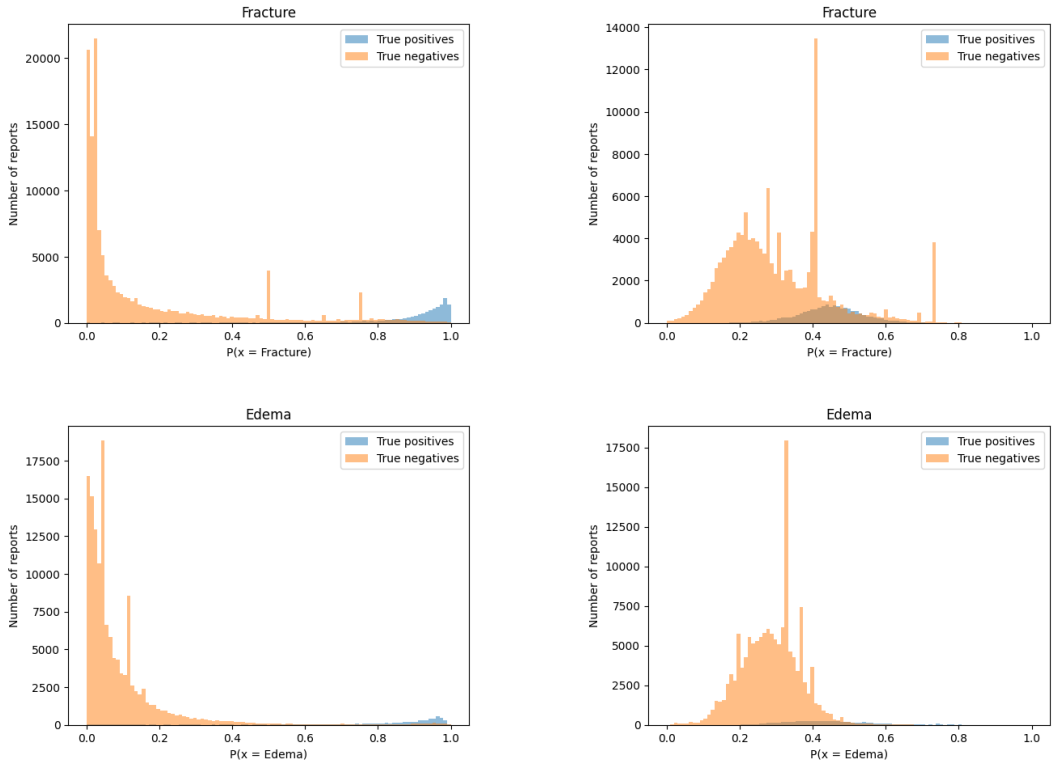

Figure 3: Histograms showing prediction distributions for "fracture" and "edema".

5.8 Disposition prediction

Performance of multilayer perceptron

Label	AUC	Precision	Recall	F1 score
Admit to Inpatient	0.68	0.63	0.57	0.60
Discharge	0.69	0.62	0.74	0.67
Place in Observation	0.48	0.05	0.00	0.00
Place in Observation - CDU	0.57	0.00	0.00	0.00

Table 4: Disposition prediction metrics for best MLP model from hyperparameter search

The overall accuracy of the best model is 0.623. We observe in table 4 that for the rare classes "Place in Observation" and "Place in Observation - CDU", the model has very weak performance. In fact, AUC=0.48 indicates that the model is slightly worse than random guessing for "Place in Observation". However, the overall accuracy and AUC's are decent compared to BERT predictors. In terms of accuracy, the MLP model outperforms all the fine-tuned transformer classifiers from Method 1 (see Figure 1), and is only slightly worse than the best model from Method 2.

Performance of BERT implementations

The following table shows the results for each of the BERT models in methods 1 and 3. The left half shows the results of method 1, the models trained only on text reports. The accuracy across all models in method 1 was between 0.57 to 0.62.

The right half of Table 5 shows the results of method 3, models trained on text reports together with the result of the zero-shot labeler. Accuracy results vary between 0.59 to 0.63.

Model	Method 1						Method 3					
	Accuracy	Label	AUC	Precision	Recall	F1-score	Accuracy	Label	AUC	Precision	Recall	F1-score
BERT	0.566	Admit to Inpatient	0.64	0.62	0.35	0.45	0.586	Admit to Inpatient	0.66	0.60	0.49	0.54
		Discharge	0.64	0.55	0.83	0.66		Discharge	0.66	0.58	0.75	0.65
		Place in Observation	0.53	0.00	0.00	0.00		Place in Observation	0.53	0.00	0.00	0.00
		Place in Observation-CDU	0.50	0.00	0.00	0.00		Place in Observation-CDU	0.53	0.00	0.00	0.00
BIO BERT	0.616	Admit to Inpatient	0.70	0.68	0.45	0.54	0.623	Admit to Inpatient	0.70	0.62	0.60	0.61
		Discharge	0.70	0.59	0.84	0.69		Discharge	0.71	0.62	0.72	0.67
		Place in Observation	0.52	0.00	0.00	0.00		Place in Observation	0.54	0.00	0.00	0.00
		Place in Observation-CDU	0.59	0.00	0.00	0.00		Place in Observation-CDU	0.56	0.00	0.00	0.00
ClinicalBERT	0.608	Admit to Inpatient	0.68	0.60	0.60	0.60	0.627	Admit to Inpatient	0.70	0.63	0.61	0.62
		Discharge	0.68	0.61	0.69	0.66		Discharge	0.71	0.63	0.72	0.67
		Place in Observation	0.53	0.00	0.00	0.00		Place in Observation	0.56	0.00	0.00	0.00
		Place in Observation-CDU	0.55	0.00	0.00	0.00		Place in Observation-CDU	0.63	0.00	0.00	0.00

Table 5: Accuracy, precision, recall and F1-score of all models. The AUC, precision, recall and F1-scores are reported separately for each classes.

Overall, ClinicalBERT in method 3 achieved the best performance by accuracy and AUC. For both methods, rare classes like Observation and Observation-CDU have much weaker performance than other classes (see Figure 4). Our findings shows that BERT implementations focused on biomedical terminology performed better than general BERT on patient's disposition prediction.

6 Analysis

6.1 Unsupervised label extraction

We observe that using pretrained NLI models for zero-shot extraction of medical label is quite effective. The CheXbert model from prior work was required several rounds of training: first pretraining BERT on medical corpus, then fine-tuning on rule-based CheXpert labels, and finally fine-tuning a small human-labeled dataset (Smit et al., 2020). In our work, we show that using a large pretrained language model for natural language inference on medical text, even without any fine-tuning, can achieve surprisingly good results (as high as 0.93 AUC for some classes).

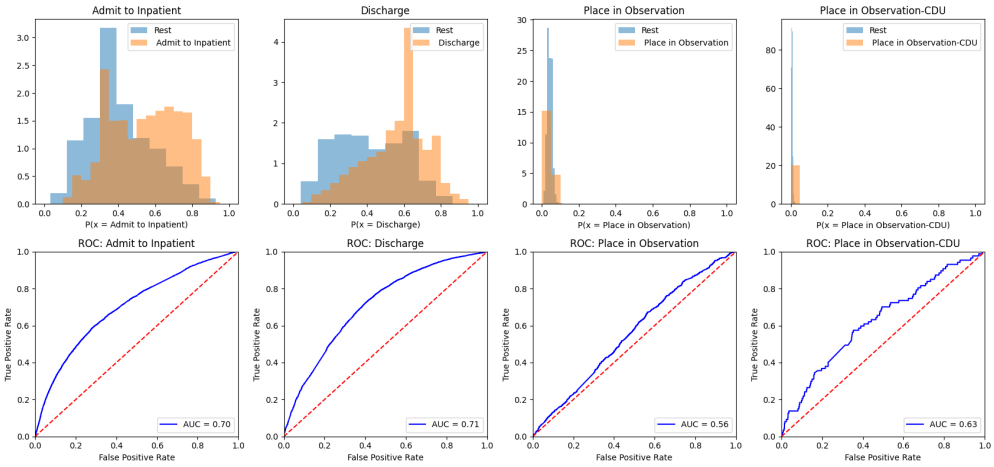

Figure 4: Histograms and ROC curves of each disposition class. Classification results after fine-tuning pretrained model ClinicalBERT.

6.2 Disposition prediction

One notable characteristic of the disposition prediction models is their inability to handle class imbalance. For common classes, the models have some non-trivial predictive ability, performing significantly better than random guessing. The two rare classes comprise roughly 5% and 0.6% of the dataset respectively, and the model effectively ignores these classes, with close to 0 precision and recall. The is true for both BERT and multilayer perceptron models. This is unsurprising, since we did not use any techniques to specifically address class imbalance.

Another notable result is that while the multilayer perceptron model underperforms the best fine-tuned BERT classifier, the gap is not large. This is somewhat surprising, since the input to the MLP network is a 39-dimension probability vector for hand-picked labels, which we assume to be significantly less informative compared to the full report text provided as input to BERT. One possible explanation is that the labels vector is a highly informative dense representation of the report text, functioning like a hand-crafted embedding vector. However, a more likely explanation is that it is inherently difficult to predict emergency department disposition using radiology reports. We note that in practice, physicians determine disposition not only using radiology reports, but also a variety of other inputs such as the radiology images, and vital signs like body temperature, heart rate, etc. This implies that a simple classifier model could be sufficient to learn the weak predictive signals in radiology reports.

7 Conclusion

Our study has several limitations. Due to memory limitations of the hardware accelerator, only a few epochs could be used for training – we believe that the performance of both label extraction and disposition predition can be improved by simply training for longer duration and with more thorough hyperparameter search. Also, the ground truth labels for the medical conditions use the results from cheXbert, which performs at roughly 80% overall accuracy. While the models have weaker performance in rare classes, future approach should use techniques to improve the performance on rare classes. Finally, in the present analysis we wanted to use only text data, but in the real-world, the radiology report would be only one of many inputs to a disposition prediction system.

In this study, we propose a method for predict patients' disposition. We have found that both BERT implementations trained on documents from biomedical domain BIOBERT and ClinicalBERT outperforms the regular BERT. The best overall method is extracting labels using the pre-trained NLI model, and then fine-tuning ClinicalBERT. However, prediction disposition based on just labels also yields decent results. As expected, rare classes have much weaker performance. We also find that augmenting the disposition classifier input with zero-shot labels is effective in improving performance of disposition classifiers. Finally, we demonstrate that using pretrained NLI models for zero-shot label extraction performs surprisingly well, especially considering that this approach is significantly

easier than prior methods that involve fine-tuning over domain-specific datasets. Given that report labeling is a widely applicable to medical fields, we would be interested to see broader adoption of this zero-shot labeling approach.

8 Acknowledgments

We like to thank Abhinav Garg for guiding us through this project and giving us feedback. Furthermore, we would like to thank Prof. David Kim, for trusting us and giving us access to the data.

References

L. Shen et al. A.E. Johnson, T.J. Pollard. 2016. Mimic-iii, a freely accessible critical care database.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison.

R. Ranganath K. Huang, J. Altosaar. 2019. Clinicalbert: modeling clinical notes and predicting hospital readmission.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Tim Schopf, Daniel Braun, and Florian Matthes. 2021. Lbl2vec: An embedding-based approach for unsupervised document retrieval on predefined topics. In *Proceedings of the 17th International Conference on Web Information Systems and Technologies - WEBIST,*, pages 124–132. INSTICC, SciTePress.

Tim Schopf, Daniel Braun, and Florian Matthes. 2023a. Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *2022 6th International Conference on Natural Language Processing and Information Retrieval (NLPIR)*, NLPIR 2022, New York, NY, USA. Association for Computing Machinery.

Tim Schopf, Daniel Braun, and Florian Matthes. 2023b. Semantic label representations with lbl2vec: A similarity-based approach for unsupervised text classification. In *Web Information Systems and Technologies*, pages 59–73, Cham. Springer International Publishing.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

A Appendix

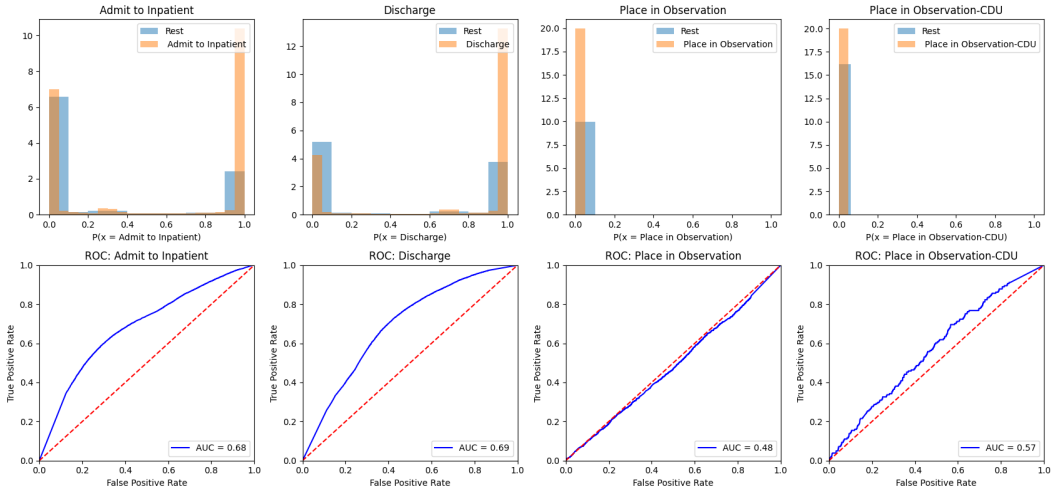

Figure 5: Histograms and ROC curves of each disposition class. Classification results from multilayer perceptron model trained on the 39 extracted labels. Like in other experiments, the model has useful ROC for the first two common classes, while basically ignoring the rare "Place in Observation" classes

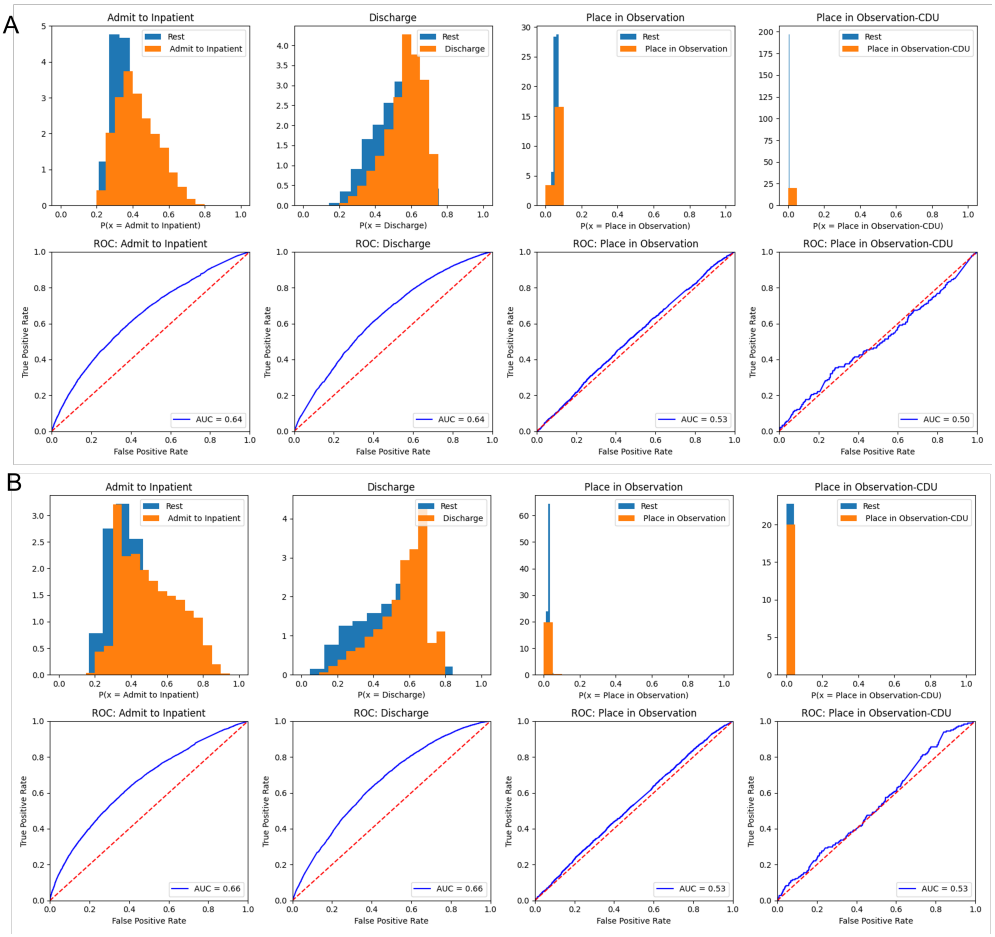

Figure 6: Histograms and ROC curves of each disposition class. Classification results after fine-tuning pretrained model (A) BERT in method 1, trained only on text reports (B) BERT in method 3, trained on text reports and result of the zero-shot labeler.

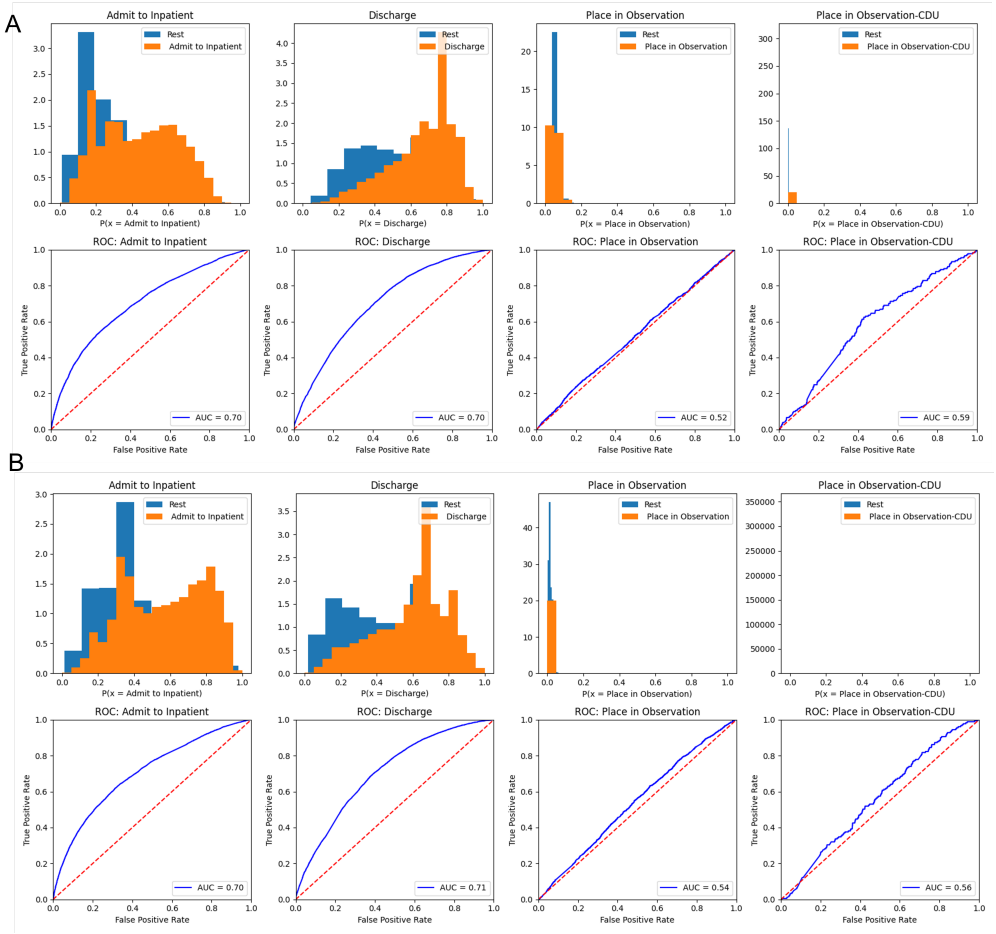

Figure 7: Histograms and ROC curves of each disposition class. Classification results after fine-tuning pretrained model (A) BIO BERT in method 1, trained only on text reports (B) BIO BERT in method 3, trained on text reports and result of the zero-shot labeler.

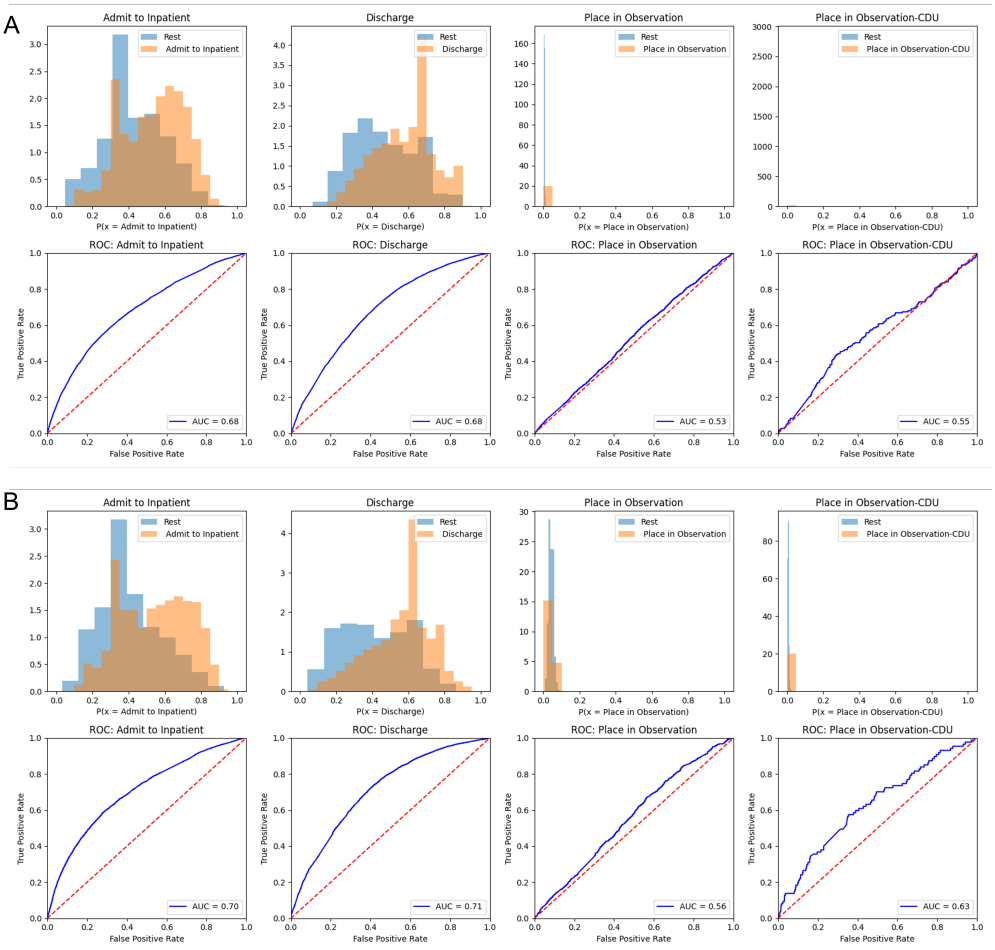

Figure 8: Histograms and ROC curves of each disposition class. Classification results after fine-tuning pretrained model (A) ClinicalBERT in method 1, trained only on text reports (B) ClinicalBERT in method 3, trained on text reports and result of the zero-shot labeler.