Examining the Efficacy of Transformer Models in Radiology Report Labeling within a Thai Hospital

1st Worranittha Larpkiattaworn
Institute of Field Robotics
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
worranittha.larp@mail.kmutt.ac.th

2nd Tretap Promwisat
Perceptra Co., Ltd.
Bangkok, Thailand
tretap@perceptra.tech

3rd Isarun Chamveha
Perceptra Co., Ltd.
Bangkok, Thailand
isarun@perceptra.tech

4th Warasinee Chaisangmongkon
Institute of Field Robotics
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
warasinee.cha@kmutt.ac.th

***Abstract*—This study explores the application of Transformer models, specifically CheXbert and CharacterBERT, in extracting labels from radiology reports in a real-world clinical setting of a Thai hospital. This setting presents unique challenges, such as spelling errors, grammar mistakes, and diverse report formats, leading to 'noisy labels'. Previous natural language processing systems, including rule-based algorithms and Transformers, have been used for this task, but they face difficulties in such environment. Despite these challenges, our research demonstrates that training Transformers on a small dataset is sufficient to outperform rule-based labelers. The study also reveals that increasing dataset size and data augmentation do not necessarily enhance accuracy, due to the potential increase in noise. Further, a comparison between CharacterBERT and CheXbert is made, showing that despite CharacterBERT's ability to handle misspellings, its accuracy does not consistently surpass that of CheXbert. The paper concludes with a case study demonstrating how CheXbert, in collaboration with rule-based labelers, can assist in identifying and rectifying potentially noisy reports, thereby aiding in label purification.**

***Index Terms*—chest X-ray report labeling, natural language processing**

I. Introduction

The task of extracting relevant labels from radiology text reports is crucial for the extensive large-scale training of medical imaging models [7]. A multitude of natural language processing (NLP) systems have been developed to facilitate this extraction process. These range from rule-based algorithms leveraging complex feature engineering rooted in medical domain knowledge [5], to Transformers which show better performance in end-to-end radiology report labeling [3], [6].

Recent advancements suggest that CharacterBERT [4] may offer a more effective solution for radiology report labeling. CharacterBERT prevents biases that may result from using a predetermined wordpiece vocabulary by using the characters of each token to construct word-level representations. This yields improved performance and robustness to noise and misspellings.

High-quality labels demand expert knowledge and can often be expensive, leading to the adoption of more economical data gathering methods like crowdsourcing and automatic annotation procedures. However, these cost-saving techniques often lead to labeling errors, creating 'noisy labels'. It has been shown that while a transformer model such as BERT is robust against artificially injected noise, it is vulnerable to noise generated from weak supervision, presenting a significant challenge in adopting BERT in real clinical settings [8].

This study explores the application of Transformer models, CheXbert and CharacterBERT, in real-world clinical scenarios within a Thai hospital. These environments present unique challenges, such as reports written by non-native speakers leading to increased ambiguity from grammatical errors or misspellings, inconsistent terminology, varying report structures, and discrepancies in diagnostic criteria, all of which contribute to heightened label noise. Despite these challenges, our research demonstrates that training CheXbert and CharacterBERT, on a realistically noisy set of reports is not just feasible, but often outperforms a rule-based labeler. Furthermore, our study underscores that larger datasets do not necessarily guarantee higher accuracy, as they could contain more noise that potentially undermines the performance. We also compare CheXbert and CharacterBERT, highlighting that although CharacterBERT can handle misspellings, its accuracy does not always surpass that of CheXbert. An evaluation of data augmentation through backtranslation reveals that it does not enhance the model's performance. Given these findings, we introduce a case study demonstrating how the CheXbert model can work in tandem with rule-based labelers to identify and rectify potentially noisy reports, thereby aiding in label purification.

II. Related Work

Several natural language processing tools have been developed to identify the disease state mentioned in chest X-ray reports. CheXpert labeler [5], developed on the largest datasets

(CheXpert and MIMIC-CXR), is a prominent natural language processing tool used to identify disease states in chest X-ray reports. It employs predefined rules based on keywords and sentence patterns to determine if the disease state is negative, positive, or uncertain.

Since rule-based labelers like CheXpert are rigid and prone to making mistakes, there is a growing interest in Transformer models that can learn contextual relationships from abundant data, offering a more adaptable solution for medical report analysis.

Smit et al. introduced the CheXbert model [6], a Transformer model for labeling 14 clinical observations from radiology reports. CheXbert is based on a BERT or Bidirectional Encoder Representations from Transformers model, trained on the same chest X-ray dataset used in CheXpert study, labeled by both CheXpert labeler and radiologists and augmented with backtranslation method. The model architecture comprises wordpiece tokenizer, a BERT model [2], and linear heads (classifiers), each for labeling different clinical observations. This model demonstrates an improvement over the previous rule-based state-of-the-art (SOTA), CheXpert labeler, with faster inference speed. However, it does have a limitation due to the wordpiece tokenizer's sensitivity to noisy words that are out-of-vocabulary or misspelled. Since the tokenizer relies on a vocabulary to split input words into subwords before feeding them to the BERT model, it may incorrectly tokenize noisy words, leading to a drop in the model's performance.

CharacterBERT [4] is a new variant of BERT designed to handle noisy words using character-level features and a Character-CNN module, which contains character embeddings, multiple 1D CNNs, and two Highway layers, instead of the subwords and wordpiece embeddings used in BERT. Unlike BERT, an input word in CharacterBERT is split into a sequence of characters to produce a single representation. The authors demonstrated that CharacterBERT, trained on both general domains (Wikipedia (EN) and OpenWebText) and medical domains (MIMIC-III dataset and PMC OA abstracts), outperformed the original BERT model in handling noise, particularly in medical tasks where noisy words are common.

To our knowledge, no study has yet explored the comparison and application of these language models to medical data in environments of data scarcity and label noise, nor demonstrated their use in cleaning large datasets, which is the focus of our study.

III. Methodology

A. Dataset for Model Development

In this study, we employ a dataset consisting of 29,849 pairs of chest X-ray images and corresponding reports, collected from Siriraj Hospital in Bangkok, Thailand. We select samples with chest X-ray images of individuals aged 15 years and older, then discard those with low-quality images and duplicated reports. After cleaning the data, we were left with 13,658 samples. We divided this dataset into 80% for training, 10% for validation, and 10% for testing.

B. Class Annotation

Two trained research assistants reviewed both chest X-ray images and radiologist reports and labeled each sample in the dataset as negative (indicating no disease) or positive states (indicating the present of disease) with seven clinical conditions including Cardiomegaly, Edema, Lung Opacity Group (Infiltration, Consolidation, Lung Opacity), Pleural Effusion, Atelectasis, Mass and Nodule. These 'human annotations' served as ground truth for our model training. Table I shows data distribution.

TABLE I
NUMBER OF NEGATIVE AND POSITIVE CASES FOR EACH CONDITION

Observation	Negative	Positive
Cardiomegaly	9,575	4,083
Edema	13,401	257
Lung Opacity Group	9,366	4,292
Pleural Effusion	12,729	929
Atelectasis	13,297	361
Mass	13,131	527
Nodule	12,074	1,584

C. Model Development

In this study, we compared two transformer-based models, inspired by the original CheXbert and CharacterBERT (as described in Section II), to identify disease states in chest X-ray reports from Thai hospitals.

We modified CheXbert, the first model, by adjusting the number of linear heads to seven clinical observations and converting each head into a binary classifier. Similar to the original model, input reports are tokenized into a sequence of subwords using a wordpiece tokenizer before being fed into our model.

We then created a new variant of CheXbert, named CheXcharacterBERT, by modifying the embedding layers of CheXbert. This model utilizes CharacterBERT instead of the BERT-base architecture, so wordpiece embedding is changed to the Character-CNN module. In addition, we use character-level tokenizer instead of wordpiece tokenizer to preprocess input reports into a sequence of characters before feeding to the model. This modification allows the model to generate a word representation based on its character-level features enhancing the model's robustness to noisy words. The difference of two model architectures are shown in Figure 1.

Both CheXbert and CheXcharacterBERT models are initialized with weights from the pre-trained CheXbert in most layers, except for the embedding layers of CheXcharacterBERT which were initialized with weights from pre-trained CharacterBERT. For training the models, we utilize cross-entropy loss, Adam optimization with a learning rate of $2 * 10^{-}5$ and a batch size of 18, following the approach from the original CheXbert. In addition, we utilize BERT tokenizer for tokenizing input reports before feeding to the models. CheXbert uses this tokenizer as wordpiece tokenizer while CheXcharacterBERT uses it as a basic tokenizer before

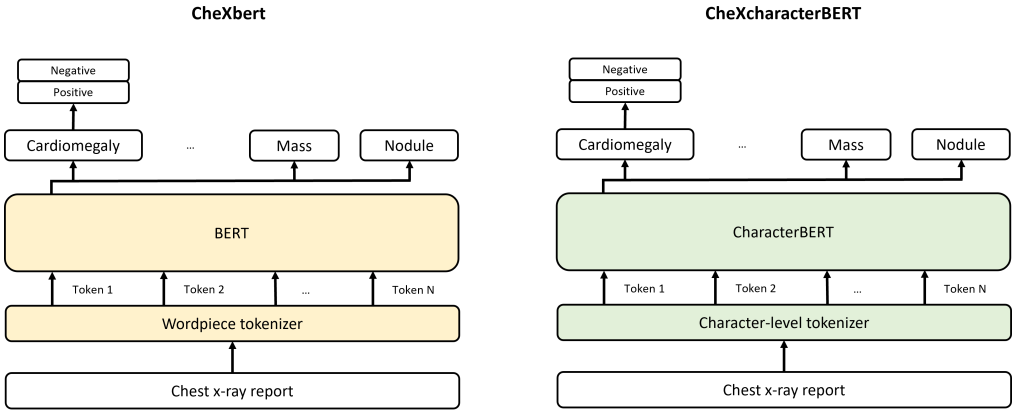

Fig. 1. Architecture of CheXbert and CheXcharacterBERT model

processing them into a sequence of characters encoded in utf-8. Both models have a maximum number of input tokens at 512.

D. Benchmark

We compare our Transformer model with a rule-based labeler, named Inspectra Labeler, which was developed on top of CheXpert [5] and described in [1]. This Inspectra Labeler modified the rule set of CheXpert Labeler with additional keywords, patterns or grammar extracted from the Siriraj dataset collected by Siriraj Hospital, Bangkok, Thailand to label seven observations in this dataset aligned with our task. The rule-based labeler comprises three stages of pipeline for labeling the reports: (1) 'mention extraction' for extracting the mentions of each observation from the Impression section in the reports that summarizes the details of chest X-ray image, (2) 'mention classification' for classifying the extracted mentions as negative, uncertain or positive by matching them with predefined rules in a rule set, and (3) 'mention aggregation' for identifying the final disease's state of each observations in the report from the states of extracted mentions. For example, in case of 'no pleural effusion or pneumothorax', this labeler extracts the mentions 'pleural effusion' and 'pneumothorax' then classifies them as negative and labels the disease's state of Pleural Effusion for this report as negative state.

E. Evaluation

We evaluated our models' classification performance using the area under the receiver operating characteristic (AUROC) curve, true positive rate (sensitivity), true negative rate (specificity), positive predictive value (PPV), negative predictive value (NPV), and accuracy. However, we did not use AUROC to compare our final model with the previous rule-based labeler, as a rule-based labeler does not predict disease states as probabilities, which are used in this metric.

IV. Results

A. Analysis of the previous rule-based labeler

Before the model experiment, we investigated the significant challenges within Inspectra Labeler, a previous rule-based labeler. The main issues come with out-of-rule keywords. For example, a rule set lacks keywords like 'LVH, LA enlargement', which can represent Cardiomegaly or keyword 'resolving of' for identifying negative disease's state. In addition, misspelling is one of the problems that the labeler struggles with, such as 'cardioemgaly' in the context of 'cardiomegaly', 'plerual efusion' in the context of 'pleural effusion', and 'atslectasis' in the context of 'atelectasis'.

To summarize, it is difficult to create a rule set to capture all diversity of chest X-ray reports with out-of-rule keywords and patterns for a rule-based labeler. In addition, the task of updating the rule set to accommodate new datasets from various hospitals with their unique report characteristics increases workload and does not ensure the long-term improvement of the labeler. Therefore, we purpose Transformer models to address these issues.

B. Model Experiment

1) Model Architecture: CharacterBERT is designed to address noisy words which are limitations of the original BERT model. We investigate the effectiveness of CharacterBERT by comparing CheXbert and CheXcharacterBERT models.

Table II shows the performance of these models. CheXbert outperforms CheXcharacterBERT specifically in identifying positive states. It has 4.3% higher sensitivity than CheXcharacterBERT. Even CheXcharacterBERT accurately labels some misspellings that CheXbert cannot such as 'cardioemgaly' in the context of 'cardiomegaly' or 'minimal fibronodular indilrearion' in the context of 'minimal fibronodular infiltration', it still performs worse than CheXbert in most normal cases. For example in Table III, CheXbert can accurately label the common pattern for indicating the negative state of the mentioned observation, 'no + observation keywords + is + noted/seen', while CheXcharacterBERT identifies it as a positive state. These results demonstrate that while CharacterBERT can handle some misspellings, it does not outperform the CheXbert model overall.

TABLE II
EVALUATION METRICS FROM OUR MODELS

Model	Average Score					
	AUROC	Sensitivity	Specificity	PPV	NPV	Accuracy
CheXbert	0.999	0.993	0.999	0.995	0.998	0.998
CheXbert-large	0.997	0.935	0.999	0.993	0.989	0.991
CheXbert-aug	0.999	0.994	0.999	0.989	0.998	0.998
CheXcharacterBERT	0.995	0.950	0.994	0.948	0.992	0.989
CheXcharacterBERT-large	0.994	0.907	0.995	0.959	0.985	0.985
CheXcharacterBERT-aug	0.994	0.940	0.993	0.934	0.992	0.988

2) Additional Dataset and Augmentation: We investigate the benefit of additional data and a data augmentation technique for improving model performance.

a) Additional Dataset: We created a larger dataset drawn from the same data repository as our main training dataset. Images and reports come from the same hospital and went through the same annotation process. This larger training set consists of 135,584 pairs of chest X-ray images and reports. After cleaning, we are left with 47,914 samples. These reports are concatenated with an existing training set and remove duplicate reports to create a 'large training set' containing 56,932 samples with data distribution shown in Table IV.

As shown in Table II, both CheXbert-large and CheXcharacterBERT-large perform slightly worse than the same models trained on the original training set, specifically in indicating positive state despite the greater number of positive samples in most conditions. Note that, due to larger size, the annotation process is laborious and tedious, and our post-training analysis suggests that the larger dataset contains more noise than the main dataset, to which BERT is vulnerable. This explains lower performance.

b) Data Augmentation Tool: We utilize backtranslation method to avoid noisy-labeled reports from additional dataset by training our models on 'augmented training set', namely CheXbert-aug and CheXcharacterBERT-aug. In this experiment, we augment the first 2,000 reports of the original training set using Google Translate to convert the reports to German and then back to English. After removing duplicates, the augmented training set contains 12,892 reports with distribution shown in Table IV.

As the results shown in Table II, the models trained on augment training set have similar scores to the same models trained on original training set in every evaluation metric which are also better than using a large training set. We observe that these models can handle more misspellings, but it still underperforms in some normal cases. For example, CheXbert-aug labels a negative pattern of Pleural Effusion, 'no active pulmonary infiltration or pleural effusion', as a positive state. This result may be due to the drawback of Google Translate that produces inaccurate reports and confuses the models. For example, 'no active pulmonary infiltration or pleural effusion is noted' is translated to 'no active lung infiltration or pleura experience is found' or 'there is RUL opacity' is translated to 'there is the ongoing opacity'. This shows that backtranslation method is not effective in improving model's performance.

TABLE IV
NUMBER OF NEGATIVE AND POSITIVE STATES OF EACH OBSERVATION IN TRAINING SET, LARGE TRAINING SET AND AUGMENTED TRAINING SET

Observation	Training Set		Large Training Set		Augmented Training Set	
	Negative	Positive	Negative	Positive	Negative	Positive
Cardiomegaly	7,636	3,290	44,424	12,508	9,006	3,886
Edema	10,714	212	56,713	219	12,638	254
Lung Opacity Group	7,501	3,425	27,044	29,888	8,858	4,034
Pleural Effusion	10,174	752	51,873	5,059	12,009	883
Atelectasis	10,634	292	55,006	1,926	12,548	344
Mass	10,513	413	54,431	2,501	12,392	500
Nodule	9,653	1,273	47,246	9,686	11,389	1,503

3) Comparison to the previous rule-based labeler: Given the results in Section IV-B1 and IV-B2, we select CheXbert as our best model due to the highest performance especially in normal cases and compare this model to Inspectra Labeler. Table V demonstrates the performance of two models. CheXbert achieves the average sensitivity of 0.993 and specificity of 0.999 that are quite better than Inspectra Labeler without requiring any predefined rule set used in a rule-based labeler.

TABLE V
THE SENSITIVITY AND SPECIFICITY OF INSPECTRA LABELER AND CHEXBERT ON TEST SET

Observation	Sensitivity		Specificity	
	Inspectra Labeler	CheXbert	Inspectra Labeler	CheXbert
Cardiomegaly	0.975	0.985	0.995	0.999
Edema	1.000	1.000	1.000	1.000
Lung Opacity Group	0.984	0.989	0.993	0.998
Pleural Effusion	0.977	0.989	1.000	1.000
Atelectasis	0.968	1.000	1.000	0.999
Mass	1.000	1.000	1.000	1.000
Nodule	0.986	0.986	1.000	1.000
Average	0.984	0.993	0.998	0.999

We observe that CheXbert can capture out-of-rule keywords such as 'LVH or LA enlargement' in Cardiomegaly and 'resolving of' for identifying negative state as shown in Table VI. It also correctly labels some misspellings that Inspectra Labeler cannot such as 'cardiomeglaly' in the context of 'cardiomegaly', 'plerual efusion' in the context of 'pleural effusion' and 'atslectasis' in the context of 'atelectasis'. In addition, training CheXbert on 10,926 samples of training set takes around 7 minutes on one NVIDIA Tesla V100 SXM2 32GB GPU. For inference speed, CheXbert also tokenizes and labels the disease's state in approximately 0.002 seconds per report faster than Inspectra Labeler that uses around 0.25 seconds. Therefore, our best model can address the limitations of a rule-based labeler with an improvement of inference speed that reduces both workload and time. However, it is important to mention that this model still struggles with some misspelled words and has a limitation of the maximum number of input tokens.

TABLE III
EXAMPLE REPORTS IN NORMAL CASES WHERE CHEXBERT OUTPERFORMS CHEXCHARACTERBERT

Observation	Example Report	Actual Label	CheXbert's Label	CheXcharacterBERT's Label
Cardiomegaly	CHEST, PA UPRIGHT No definite pulmonary infiltration or consolidation. **No cardiomegaly is noted.** Both costophrenic angles are clear.	0	0	1
Lung Opacity Group	CXR (PA) Cardiomegaly with mild increase pulmonary vasculature. The fibrotic band is seen at RLL. **No active pulmonary infiltration.** No pleural effusion.	0	0	1

TABLE VI
EXAMPLE REPORTS WHERE CHEXBERT OUTPERFORMS INSPECTRA LABELER

Observation	Example Report	Actual Label	CheXbert's Label	Inspectra Labeler's Label
Cardiomegaly	CHEST : P.A. view . No active pulmonary infiltration is noted . **LVH is still observed .** Others are not remarkable .	1	1	0
Lung Opacity Group	Findings: Normal heart size and pulmonary vasculature Normal hili. **Resolving of interstitial opacity in LLL.** Unchanged of plate atelectasis in LLL.	0	0	1
Pleural Effusion	Chest PA **Decreased of bilateral plerual efusion** from 06/01/2015. Interstitial thickenng along both lungs. Prominent heart size.	1	1	0

C. Case Study

We demonstrate how a Transformer model can be used in conjunction with a rule-based labeler to clean larger repositories. To this end, we use the whole chest X-ray data repository collected from Siriraj Hospital. This dataset comprises 884,753 pairs of chest X-ray images and their corresponding radiology reports. We select samples following the same criteria as the main dataset, resulting in 360,603 samples.

We labeled this repository using both Inspectra Labeler and our CheXbert model and compared the results. The reports fell into two groups: (1) 10,795 reports had differing labels between our model and Inspectra Labeler, and (2) 349,808 reports had consistent labels from both labelers. This approach significantly reduced the review workload, as human annotators only had to review 3% of the repository.

We select 100 samples in each condition from the first group ranked by the difference between CheXbert's predicted probabilities and Inspectra Labeler's results from highest to lowest to investigate in more detail. Compared to Inspectra Labeler, CheXbert can capture more keywords and patterns, resulting in higher number of accurate labels, as shown in Table VII. However, there are some incorrect labels due to new keywords and patterns such as 'pneumohydrothorax' in Pleural Effusion or 'much clearing of + observation keyword' which are not present in our model's training set, so these reports can be re-annotated and add to the training set to improve dataset accuracy.

In another group where both Inspectra Labeler and CheXbert yield consistent labels, we considered the possibility of incorrect labels by both models. To enhance the quality of this group, we applied an in-house chest X-ray image classifier to identify abnormalities [1]. Our primary focus was on cases where CheXbert's predictions differed from those of the image classifier. In these cases, we observed that the reports often included vague terms such as 'mild,' 'small,' or 'decrease.' While our CheXbert model labeled these reports as positive, the image classifier, due to the absence of apparent abnormal features in the X-ray images, classified these samples as negative. We can then review and re-annotate this subgroup.

Overall, the iterative approach of comparing model-generated labels, identifying discrepancies, refining the models and the labeling process, and updating the models as they improve allows for increased dataset accuracy and reduced labeling errors, contributing to a more reliable and valuable repository for future research and applications.

TABLE VII
NUMBER OF ACCURATE LABELS FROM CHEXBERT AND INSPECTRA LABELER IN THE FIRST 100 SAMPLES OF EACH CONDITION

Observation	CheXbert	Inspectra Labeler
Cardiomegaly	97	3
Edema	71	29
Lung Opacity Group	98	2
Pleural Effusion	88	12
Atelectasis	91	9
Mass	87	13
Nodule	81	19

V. Conclusion

In this study, we present transformer-based models for identifying the disease's state on chest X-ray reports from a Thai hospital that improve over a previous labeler, coupled with a case study of using our model with a rule-based labeler to identify and extract noisy-labeled reports.

A previous rule-based labeler faces the challenges in capturing out-of-rule keywords and patterns in chest X-ray reports. To address these problems, we demonstrate CheXbert model that can achieve the average sensitivity of 0.993 and specificity of 0.999 in identifying the disease's state from Thai chest X-ray reports and has faster inference speed compared to a rule-based labeler without using addition information such as a predefined rule set. In addition, our final model can be used with a rule-based labeler to identify and extract noisy-labeled reports from the dataset with inaccurate labels for improving dataset accuracy which reduce both workload and time of radiologists to assess all data. Given the utilization of reports written by non-native speakers in this study, our model may also be applicable for those who are non-native speakers but communicate in English in other regions.

In addition, our study also demonstrates the effectiveness of CharacterBERT in handling some misspellings and the advantage of backtranslation method with Google Translate for data augmentation. However, CharacterBERT does not outperform the BERT model in overall performance and Google Translate can produce some errors in translating chest X-ray reports which lead to model confusion.

However, our model still has the downside in capturing some misspelled words in chest X-ray reports and a limitation of the maximum number of input tokens. Due to a small test set from a single hospital, there are also few examples that our model outperforms a previous rule-based labeler, and the model is not evaluated for generalizability. Additional datasets across other Thai hospitals with different characteristics are required to investigate the model's generalizability in future work. In addition, it would be worth to verify other results of our model on a large dataset, as discussed Section IV-C, to ensure data accuracy and gain more insights. This verification process, performed by trained research assistants, may be time-consuming and costly, but it would be feasible if resources were not a limitation.

Acknowledgment

We would like to express our gratitude to the Department of Radiology, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand, for their support in providing the chest X-ray images and reports dataset. Additionally, we thank Perceptra Co., Ltd. for contribution of the Inspectra Labeler and Inspectra model. Approval of all ethical and experimental procedures for this work was granted by KMUTT-IRB (Cert. No. KMUTT-IRB-COE-2021-043) performed in line with the international guidelines of human research protection.

References

[1] Isarun Chamveha, Trongtum Tongdee, Pairash Saiviroonporn, and Warasinee Chaisangmongkon. Local adaptation improves accuracy of deep learning model for automated x-ray thoracic disease detection: A thai study. *arXiv preprint arXiv:2004.10975*, 2020.

[2] 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, 2019.

[3] Ignat Drozdov, Daniel Forbes, Benjamin Szubert, Mark Hall, Chris Carlin, and David J Lowe. Supervised and unsupervised language modelling in chest x-ray radiological reports. *Plos one*, 15(3):e0229963, 2020.

[4] Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, 2020.

[5] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

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

[7] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[8] Dawei Zhu, Michael A Hedderich, Fangzhou Zhai, David Ifeoluwa Adelani, and Dietrich Klakow. Is bert robust to label noise? a study on learning with noisy labels in text classification. *arXiv preprint arXiv:2204.09371*, 2022.