DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Contrastive Pre-Training For Radiology Reports

Ilayda Ezgi Zengin

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Contrastive Pre-Training For Radiology Reports

Contrastive Pre-Training für Radiologieberichte

Author:	Ilayda Ezgi Zengin
Supervisor:	Prof. Dr. Daniel Rückert
Advisor:	MSc. Philip Müller
Submission Date:	15.05.2023

I confirm that this master's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 15.05.2023

Ilayda Ezgi Zengin

Abstract

Working on biomedical documents is becoming more and more significant with the increased volume of biomedical literature followed with growing interest among researchers. Articles, journals, records and clinical documents from biomedical domain not only contain new discoveries and insights but also create an overwhelming demand and need for various tasks such as information extraction in biomedical research. In recent years transformer-based language models have proven quite successful in the field of Natural Language Processing (NLP). These models require huge amounts of training data and are therefore typically pre-trained on unlabelled datasets using self-supervised objectives like Masked Language Modeling (MLM) as proposed in BERT model. While there are also models pre-trained in biomedical domain, the pre-training objectives do not utilise the structure of medical documents enough.

In this thesis, an approach that leverages the semi-structured nature of radiology reports with the application of contrastive methods on the sections of the reports is followed. Additionally, by applying several data augmentation strategies on sentence level, this thesis aims to construct a permutation invariant pipeline for the pre-training of radiology reports. According to experiment results, models pre-trained in this thesis achieves comparable performance with the state of the art models in biomedical NLP for four downstream tasks. With the combination of different data augmentation strategies and contrastive learning in pre-training pipeline, experiment results propose different directions for further research.

Kurzfassung

Mit dem immer mehr an Bedeutung gewinnt die Arbeit an biomedizinischen Textdaten, ein zunehmendes Volumen an biomedizinischer Literatur, die mit wachsendem Interesse unter ihnen Forscher folgte. Aufzeichnungen und klinische Dokumente aus dem biomedizinischen Bereich enthalten nicht nur neue Entdeckungen und Einsichten, sondern schaffen auch eine überwältigende Nachfrage für verschiedene Aufgaben wie die Informationsextraktion in der biomedizinischen Forschung. Transformer basierte Sprachmodelle haben sich in den letzten Jahren im NLP als recht erfolgreich erwiesen. Diese Modelle erfordern riesige Mengen an Trainingsdaten und werden daher in der Regel auf unbeschrifteten Datensätzen vortrainiert, wobei selbst überwachte Ziele wie MLM verwendet werden, wie im BERT-Modell vorgeschlagen. Es gibt zwar auch Modelle, die im biomedizinischen Bereich vortrainiert sind, aber die Vortrainingsziele nutzen die Struktur medizinischer Dokumente nicht ausreichend. In dieser Arbeit wird ein Ansatz verfolgt, der die halb-strukturierte Natur von Radiologieberichten mit contrastive Pre-training auf die Abschnitte der Berichte nutzt. Darüber hinaus zielt diese Arbeit darauf ab, durch die Anwendung mehrerer Datenaugmentation-Strategien auf Satzebene eine Permutation invariante Pipeline für Pre-training von Radiologie Berichten zu konstruieren. Den Versuchsergebnissen zufolge erreichen die in dieser Arbeit vortrainierten Modelle für vier nachgelagerte Aufgaben eine vergleichbare Leistung wie die State-of-the-Art-Modelle in der biomedizinische NLP. Mit der Kombination verschiedener Datenerweiterungsstrategien und contrastive Pre-training pipeline schlagen die Experimentalergebnisse unterschiedliche Richtungen für die weitere Forschung vor.

Acronyms

CV Computer Vision. iv, 9, 12

CXR Chest X-Ray. iv, 2, 8, 10, 11, 15, 17, 19, 20, 24–26, 31, 33, 34, 38, 39

DICOM Digital Imaging and Communications in Medicine. iv, 15

EHR Electronic Health Record. iv, 15

GANs Generative Adversarial Networks. iv, 10

GPT Generative Pre-trained Transformer. iv, 10

IOB2 Inside-Outside-Beginning. iv, 26

IVF The Inverted File Index. iv, 21

MLM Masked Language Modeling. iii, iv, 2, 10, 11

NER Named Entity Recognition. iv, 3, 4, 6, 12, 27, 34, 39, 42, 45

NLI Natural Language Inference. iv, 3, 17, 25, 32–35, 39

NLP Natural Language Processing. iii, iv, 1–4, 9, 12, 13, 15, 38–40

PHI Protected Health Information. iv, 15

RE Relation Extraction. iv, 4, 6

RSM Radiology Section Matching. iv, 11, 12, 19, 22, 23, 33, 38, 41

VAEs Variational Autoencoders. iv, 10

VLP Vision Language Processing. iv, 2, 35

Contents

1. Introduction

1.1. Motivation

The development of transformer based language models in recent years have significantly improved the performance of many NLP tasks in different domains [1]. However these models require huge amount of training data and perform poorly for domain specific corpora such as biomedical domain vocabulary. Biomedical text with its complex and specific linguistic characteristics causes additional challenges compared to general domain [26]. Therefore, adaptations of transformers based language models on biomedical data plays a crucial role to improve the clinical research while providing valuable information about the data.

It is difficult to handle all challenges that occurs with domain specific corpora. To deal with some of them, many transformer based language models have specialized on domain specific data and they have developed various solutions by focusing on different challenges. Some approaches addressed these issues by pre-training a model from scratch with a new vocabulary that is constructed from domain specific data as in SciBERT [3]. Although focusing on the linguistic characteristics for domain specific vocabulary can provide significant improvements on learning contextual information, it also requires high computational resources and huge amounts of training data when the model is pre-trained from scratch. As an alternative, the existing pre-trained language models are used as the initial model and are further trained to learn domain specific vocabulary as in BioBERT [26, 46].

Even though training the model from scratch with domain specific vocabulary creates high quality representations and embeddings for the specific domain, it is highly costly in terms of computational resources. On the other hand, using the existing pre-trained language models saves from computational resources while being more environment friendly [24, 10]. Additionally, using existing pre-trained models provides the opportunity to invest in the ability of the model's learning in terms of training objectives and to focus on other challenges that transformer based language models can benefit such as model architecture [11]. Even though transformers based language models achieve high performance and and are in line with expert-level performance, these systems still require huge amount of domain specific train data. Therefore, learning the domain specific vocabulary in unlabeled data setting for biomedical domain requires special attention and further research.

1.2. Problem Statement and Objectives

Along with the development of language models in biomedical domain, new objectives to learn the semi-structure of the biomedical documents have been introduced in biomedical NLP. Different than the objectives of transformer based language models such as in BioBERT [26] that treat each sentence as an independent sentence, new objectives such as contrastive learning aims to leverage the semi structure of the biomedical documents which has been used in CXR-BERT [4]. This thesis focuses on developing a language model for biomedical documents, specifically for radiology reports by using contrastive learning on the sections of the reports in the training process. It also aims to experiment and investigate the effect of different data augmentation strategies on sentence level within four downstream tasks.

1.3. Contributions

As described in section 1.2, main goal of this thesis is developing a language model for radiology reports with a contrastive learning approach. The model that inspired this thesis, CXR-BERT model [4], uses contrastive learning as a novel pre-training objective in addition to MLM objective by focusing on semi-structured characteristics of the radiology reports. CXR-BERT also proposes an effective self- supervised Vision Language Processing (VLP) approach for paired image and text data in biomedical domain. To this end, adaptation of transformer based language models with different approaches has been introduced to improve the performance of the models on biomedical documents. Compared to existing biomedical and clinical transformer based language models, CXR-BERT applies sentence level data augmentation with sentence shuffling within *Findings* and *Impression* sections of the reports in the pre-training phase.

This thesis extends these approaches and objectives introduced by biomedical and clinical language models, including CXR-BERT and presents the following contributions:

- By utilizing the semi-structured characteristics of the radiology reports, this work presents four different data augmentation strategies on sentence level in the pre-training phase of the radiology reports.
- The thesis introduces five uncased and four cased Chest X-Ray (CXR) domain language models. Each model is pre-trained with a different data augmentation

strategy on sentence level and models follow the contrastive learning objective introduced in CXR-BERT for the pre-training phase of the radiology reports.

- In this thesis, comprehensive experiments are conducted to evaluate the effect of augmentation strategies and performance of the models in Multi Label Classification, Named Entity Recognition (NER), Zero-Shot Classification and Natural Language Inference (NLI) downstream tasks with different biomedical domain datasets. Downstream tasks introduce varying evaluation strategies for results.
- The pre-training approach in this thesis considerably improve the Multi Label Classification performance of CheXpert labels through novel training procedure that leverages contrastive learning and data augmentation in a cost effective resource setting.

1.4. Outline

Following the introduction chapter above, the rest of the thesis will provide an overview about the background and related work information in Chapter 2. This chapter aims to explain fundamental concepts and approaches introduced in relevant studies to clarify the methodology throughout the thesis. After going through the essential concepts and related studies behind this thesis, Chapter 3 includes an extensive description of the data that has been used in pre-training and in downstream tasks. Chapter 4 consists of detailed explanation about the majority of work done as methodology and implementation. This chapter dives into the training pipeline, data augmentation, model architecture, hyperparameter settings in pre-training and implementation choices that have been made throughout the thesis to achieve the described objectives.

Following Chapter 4 about methodologies, Chapter 5 consists of the explanations of the downstream tasks that pre-trained models have been evaluated while providing information about the hyperparameter configurations of the tasks. Based on the information gained from the previous chapters, Chapter 6 presents the evaluation methodologies and associated results for the corresponding downstream tasks. Challenges encountered during the downstream task implementation and adapted evaluation choices are also discussed in detail and are included in Chapter 6.

As the final chapter, Chapter 7 summarizes key points addressed throughout the thesis and all work done to fulfill the stated objectives. This chapter is finalized with pointing out the possible future work that can be conducted in biomedical NLP.

2. Related Work and Background

This chapter aims to provide detailed information of relevant research associated with the work presented in this thesis for a better understanding. Following the overview of related work used in the thesis, it describes the background information needed in the pre-training process with a deeper insight.

2.1. Processing Biomedical Documents

Learning the context and characteristics of biomedical text data with its complex structure in terms of semantics creates challenges for language models that are trained with general domain data. Previous works have shown that using the language models trained in general domain such as BERT [9] on biomedical text data gives unsatisfactory results [26]. One of the main reasons for the poor performance of these models is the significant difference in the word distributions between general domain corpora and biomedical domain corpora and the need for domain adaptation. Even though transformers based language models have achieved strong results and contributed effectively on many NLP tasks, the difference between biomedical and general domain vocabulary limits their performance [26]. This problem has led to two possible approaches to provide acceptable and satisfactory results [13].

First approach is performing additional training with biomedical data on top of a base model as the initial model, which is the continuous pre-training approach for biomedical data [26]. While pre-trained vocabulary of the baseline model remains unchanged, the model weights are adapted to the new domain. The first biomedical domain specific and BERT based model is BioBERT which was pre-trained on PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC) as biomedical domain data for 23 days. Compared to BERT, BioBERT showed significant performance increase in terms of F1 score for the downstream tasks such as biomedical NER and biomedical Relation Extraction (RE). Beside achieving state of the art performance in various biomedical domain tasks, BioBERT did not perform major changes in the model architecture or in pre-training objectives and kept the the same model structure as BERT [26]. With minimal task-related modifications, BioBERT outperformed previous models for biomedical domain tasks and addressed the growing need for language models specific to biomedical domain.

On the other hand, continual pre-training of general domain pre-trained models with domain specific data as in BioBERT, creates a vocabulary disadvantage since the adapted model vocabulary is the same as the vocabulary of the baseline model and therefore do not represent the target biomedical domain [13]. To overcome this disadvantage, researchers conducted pre-training of the models from scratch with domain specific data which is the second approach to solve domain adaptation problem. In this approach, the baseline model is initialized from scratch and is pre-trained only with domain specific data such as biomedical data. A major advantage of this approach is having a domain specific vocabulary. While models with continual pre-training approach do not contain biomedical terms as words and divide these terms into subword pieces, models pre-trained from scratch have these terms in their vocabulary as single words [13, 16, 46]. Another advantage of pre-training from scratch only with domain specific data is the optimization of the model weights only for the specific domain. In continual pre-training approach, models trained with general domain data need to adapt the non-convex optimization of the pre-training process to domain specific data which may not be performed completely and fully [13].

Whether trained from scratch or with continual pre-training approach, training language models require huge amounts of domain specific annotated data. Since annotated data in specific domains such as biomedical is limited, objectives that reduce the dependency on annotated data as introduced in BERT [9] is used in pre-training of domain specific models [47]. Other than focusing on the self supervised pre-training objective, researchers also investigated how to effectively utilize the significant amount of text corpora from similar domains that can be useful. SciBERT is one of the models that performs continuous pre-training based on BERT and is pre-trained from scratch with large scientific corpora. The unsupervised training of SciBERT model on large amount of scientific publications not only achieved state of the art results in various general and domain specific downstream tasks but also provided insights about the effectiveness of domain specific vocabulary in the pre-training process [3, 47].

Even though SciBERT contributed to the development of scientific domain language models significantly, it also showed that the biggest disadvantage of training from scratch for domain specific data is the need for large training resources in terms of computation, time and data [16]. Additionally, decisions about the model architecture, pre-training objective, hyper-parameter settings and cost function are left open to experiment which can be time consuming and can yield to unsatisfactory results. By considering the required substantial amount of computation power and training data, researchers started to work on solutions under constrained resources for computation and data [46, 16]. Especially after the success of SciBERT with a domain specific vocabulary, researchers started to explore novel approaches to incorporate and learn domain specific vocabulary without pre-training the complete model from scratch [46].

A research that focuses on reducing required computational resources and having low cost while introducing a proper vocabulary is the exBERT model [46] and its novel training approach. ExBERT addresses the domain adaptation issue for biomedical data by adding and learning a small extension module and augments word embeddings of BERT with the new embeddings from biomedical domain vocabulary in this module. To incorporate the learnt module for biomedical domain vocabulary, authors also introduced a trainable weighted sum operation so that original BERT embeddings outputs and learnt module outputs can be combined. Additionally, they experimented with different extension vocabulary sizes to measure the impact of vocabulary size on performance. Beside being efficient by keeping the original model the same and only training a small module for extended vocabulary, exBERT also achieved higher performance than BioBERT under limited resources for NER and RE tasks in biomedical domain.

Another research that studies the effect of adapting domain-specific vocabulary under limited resources is conducted by the authors of AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain study [16]. In this research, authors considered the vocabulary as an optimizable parameter that can be expanded according to tokenization statistics as the relative importance metric. To prevent overfitting of downstream data, they introduced a regularization technique that leverages contrastive learning framework. Each input is tokenized twice; once with baseline model vocabulary and once with adapted vocabulary. While same layer of the encoder outputs of two different tokenizations are considered as positive pair, different layers of encoder outputs for two tokenizations are accepted as negative pairs. The objective of the model is learning positive pairs in pre-tranining while preventing the model from overfitting the downstream data. Without large domain specific pre-training data and only by updating the vocabulary from downstream data during fine-tuning, they provided comparable results and consistent performance improvements in various downstream tasks in diverse domains such as biomedical, computer science, news, and reviews [16].

Differences between general domain and biomedical domain in terms of syntax, grammar, abbreviations and domain related jargon lead researchers to focus on domain-specific vocabulary [47]. While researchers investigate ways to close the gap between general and biomedical domain, their novel approaches also introduce new parameters to learn and new decisions to take in terms of model architecture and pre-training objectives.

On the other hand, effective model initialization with clearly defined objectives, application of regularization to prevent overfitting on limited domain data and better generalization performance on unseen data, promote utilization of pre-trained language models [47]. By also considering the fact that pre-trained language models do not rely on labeled data, and are pre-trained on unlabeled data with different self supervised learning objectives, this thesis focuses on adopting the pre-trained language model improvements for biomedical domain, specifically for radiology reports.

2.2. Transfer Learning in Biomedical NLP

Transfer learning is a machine learning technique that utilizes a model that was pre-trained on one task to learn a second related task. Especially in domains that labeled data is scarce such as biomedical domain, transfer learning utilizes the weights of the pre-trained model on general domain as an initialization. The weights of the model are further updated fully or partially during re-training or continuous pre-training, or updated during fine tuning of the domain specific task. In this way transfer learning leverages the labeled data of pre-trained models in general domain for limited domain specific annotated data, and achieves significant improvements on domain specific downstream tasks. BioBERT model which was initialized with the weights of BERT model and continuously pre-trained with biomedical domain data is one of the examples of transfer learning approach in biomedical NLP with its achievements [53].

Application of transfer learning from general domain to biomedical domain enabled models to improve their performance compared to models pre-trained from scratch. However, even though transfer learning provides a clear objective to follow from the pre-trained model and acts as an effective start in terms of model weights, applying transfer learning approach for domains such as biomedical has the undeniable problem of not learning meaningful representations of the domain specific data. This problem occurs not only due to the differences between general and domain specific data, but also due to the initial model training objectives and domain specific tasks. In the cases that training task of the first model is not relevant for the specific domain, self-supervised learning is a more suitable approach for pre-training, since the model not only learns about biomedical domain, but also leverages the unlabeled data [53, 25].

2.3. Self Supervised Learning in Biomedical NLP

Unlike the limited labeled data in biomedical domain, unlabeled biomedical data is extensive which can be utilized with different objectives of self supervised learning to learn the complex structures and types of the unlabeled biomedical data [25]. One example for self supervised learning approach is pre-training the models with an association learning objective such that models are pre-trained to predict the association between two samples of data from the same patient. After the pre-training phase, the model can be finetuned on a biomedical downstream task with a smaller labeled dataset. The pre-training of the model in self supervised learning acts as a feature extraction and attribute learning step. In this way the model learns to extract useful attributes and features from the unlabelled data without seeing any labelled data in the pre-training [25, 47]. Two major approaches for self supervised learning are Contrastive Learning and Generative Learning.

2.3.1. Contrastive Learning

Contrastive learning is a self supervised learning approach that aims to learn meaningful representations and common attributes of data by contrasting similar and dissimilar pairs of examples. The goal is to make the representations of similar examples more related and those of dissimilar examples more unrelated, thus improving their discriminative power [11]. The main objective of the contrastive learning approach is predicting whether a pair of samples are positive pairs or negative pairs. By receiving huge amount of pairs as training data and predicting the similarity association within pairs, the model learns to extract features from the data and therefore achieves an extensive understanding about the data [25]. The objective of contrastive learning for representation learning was introduced by [34] as the InfoNCE loss where NCE stands for Noise-Contrastive Estimation [37].

As one of the key components of contrastive learning, data augmentation is widely used to generate new training samples as positive pairs for the learned representations of the training data. Data augmentation can be applied with various transformation techniques to the original data, such as rotations, translations, scaling, and color distortions for image data and synonym replacement, insertion, deletion, and swapping, and shuffling of words in textual data [4, 6, 11]. The choice of data augmentation technique depends on the specific dataset and task, and different techniques can be more effective for different types of data.

For instance, approach introduced in SimCLR [6], the model learns the representations of positive pairs by applying three augmentation strategies as random cropping, random color distortions, and random Gaussian blur to training images and by contrasting the augmented images with other examples in the same and different modalities [6].

A study that introduces the effective contrastive learning approach is proposed in [57] for medical imaging. In this study authors apply an unsupervised learning strategy to improve medical visual understanding and representation. Their framework uses contrastive learning by maximizing the agreement between medical images and their corresponding textual descriptions from MIMIC-CXR dataset. Their approach in pre-training for medical image understanding with contrastive learning provides an effective use of the data to extract information and requires no additional expert input [57]. Especially considering the fact labelled medical data is often limited and the data is sparse, their approach demonstrated the potential of contrastive learning in biomedical field [37].

In another work introduced as CLIP [37], authors propose a new approach to learn representations of images and text simultaneously via contrastive learning. The model is pre-trained to understand the relation between pairs of images and their associated textual descriptions by learning to maximize the similarity between an image and its corresponding text description and minimize the similarity of the image with randomly sampled text descriptions. In this way, model is encouraged to learn a shared and joint representation of the image and its corresponding text. The CLIP paper introduces a novel approach for pre-training the model to understand natural language and image relation with a contrastive learning approach. Additionally, it has demonstrated impressive performance on several benchmarks, and has the potential to be used for a wide range of tasks in Computer Vision (CV) and NLP [37].

Another study that leverages contrastive learning is introduced in SimCSE [11] paper. In this study, authors focus on learning sentence embeddings by using contrastive learning and they propose a novel approach to capture the semantic similarity of sentences. The model is pre-trained to differentiate the pairs of similar and dissimilar sentences by maximizing the similarity between pairs of similar sentences, and minimizing the similarity between pairs of dissimilar sentences. This objective helps the model learn a shared representation while learning the semantic meanings of the sentences. Other than proposing a simple contrastive learning objective for sentence embeddings and providing improved the state-of-the-art results for sentence similarity tasks, their approach also draw attention to data augmentation on sentence level rather than word level.

Contrastive learning has shown significant results and introduced novel approaches to improve the information gained from unlabelled domain specific data. Especially in biomedical domain in which labeled data is limited and huge amount of unlabeled data exists, the models with contrastive learning objective grasp a comprehensive understanding and meaningful representations of the data such as medical reports, scientific articles, and electronic health records. Overall, contrastive learning has shown great potential in the biomedical domain, and its application leads to significant improvements in various tasks in the future [25].

2.3.2. Generative Learning

Generative learning is a self supervised learning methodology that aims to model and encode the input data to an explicit vector and try to decode and reconstruct the input data from the explicit vector. In this way, the generative model learns the underlying distribution of the input data while generating the reconstruction of it. In other words, the objective of generative learning is to learn the distribution that can be used to generate new data instances from the same distribution as the original input data. Compared to discriminative learning that aims to learn the decision boundary between classes or labels, generative learning focuses on generating new instances that are similar to the input data [12]. Examples for generative learning approach include variational autoencoders Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). VAEs aim to model the structure of the input data by using a latent variable model to maximize the probability of the data. Since this goal is difficult to achieve directly, VAEs try to maximize a lower bound on the log likelihood of the data [23]. On the other hand, GANs have an objective that combines a generative model and discriminative model. While generative model aims to create the input data distribution, discriminative model classifies the instances as true distribution or the generative model distribution. According to the difference between distributions, generative model updates its parameters [12].

Another example for generative learning approach is autoregressive models. Autoregressive models aim to learn the distribution of a sequence of data by using the conditional distribution of each sample given the previous samples in the sequence. In this way, the model learns to generate each sample conditioned on the previous ones [38]. Generative Pre-trained Transformer (GPT) series are important examples of autoregressive language models. By leveraging the idea of encoding the joint probability in generative learning, GPT series learn to condition on long-range information from the data sequence. Achievements of the GPT-3, the latest GPT with significant results, already shows that generative language models act as a milestone and play an important role for the development of general language systems [5]. Studies such as BioBART [55], use generative learning for various biomedical domain tasks while bridging the gap between generative learning approaches and domain specific data and emphasizing the lack of biomedical generative models.

2.4. CXR-BERT

CXR-BERT [4] is a CXR domain-specific language model that utilizes contrastive learning and text augmentation in its novel pre-tranining process on radiology reports. The pre-training consists of three phases to successfully acquire the semantics and characteristics of the reports. For this purpose, authors build their custom WordPiece [52] vocabulary from PubMed Abstracts, MIMIC-III [20] clinical notes and MIMIC-CXR radiology reports in the first phase. In this way, the model learns to represent vocabulary as full words rather than breaking the words into subwords. In the second phase, authors pre-train a randomly initialized BERT model with MLM task on their corpora. To further specialize the model on CXR reports in the third phase, they continue pre-training the model with MLM task on MIMIX-CXR radiology reports and introduce a Radiology Section Matching (RSM) task as a new pre-training task. The radiology matching task aims to match the *Findings* and *Impression* sections of the MIMIC-CXR radiology reports. A radiology report consists of several sections. The *Findings* and *Impression* sections contain summaries of radiological findings and clinical assessments [4]. Details of the dataset are explained in section 3.1.

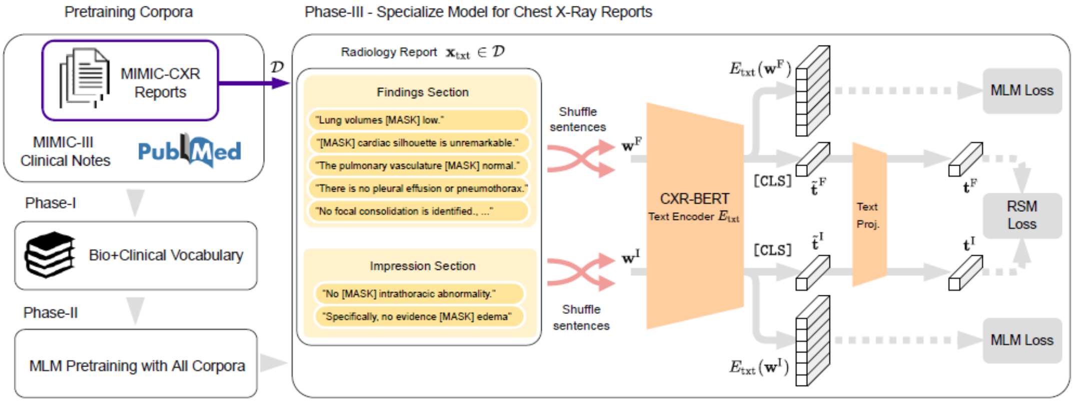

Pretraining Corpora
Phase-III - Specialize Model for Chest X-Ray Reports
MIMIC-CXR Reports
MIMIC-III Clinical Notes
PubMed
$\mathcal{D}$
Phase-I
Bio+Clinical Vocabulary
Phase-II
MLM Pretraining with All Corpora
Radiology Report $\mathbf{x}_{\text{txt}} \in \mathcal{D}$
Findings Section
"Lung volumes [MASK] low."
"[MASK] cardiac silhouette is unremarkable."
"The pulmonary vasculature [MASK] normal."
"There is no pleural effusion or pneumothorax."
"No focal consolidation is identified., ..."
Impression Section
"No [MASK] intrathoracic abnormality."
"Specifically, no evidence [MASK] edema"
Shuffle sentences
Shuffle sentences
$\mathbf{w}^{\text{F}}$
$\mathbf{w}^{\text{I}}$
CXR-BERT
Text Encoder E_{txt}
$E_{\text{txt}}(\mathbf{w}^{\text{F}})$
$E_{\text{txt}}(\mathbf{w}^{\text{I}})$
[CLS]
$\tilde{\mathbf{t}}^{\text{F}}$
[CLS]
$\tilde{\mathbf{t}}^{\text{I}}$
Text Proj.
$\mathbf{t}^{\text{F}}$
$\mathbf{t}^{\text{I}}$
MLM Loss
RSM Loss
MLM Loss

Figure 2.1.: CXR-BERT language model encoder structure with three phases of pre-training [4]

To introduce permutation invariance for *Findings* and *Impression* sections of the radiology reports, authors randomly shuffle sentences within each section as an augmentation strategy on sentence level for the pre-training of the model in the third phase. As shown in 2.1, following shuffling the sentences, the model continues pre-training for MLM task on radiology reports while starting the new training for the RSM task.

Authors leverage the section structure of the reports and introduce their contrastive loss in this section by using *Findings* and *Impression* sections from the same report as a positive pair and treating *Findings* and *Impression* sections from different reports as negative pairs. After retrieving the [CLS] token embeddings corresponding to the *Findings* and *Impression* sections of each report, the token embeddings are projected to a lower dimension with a two-layer perceptron. Denoting (t_i^F, t_i^I) as the pair of *Findings* and *Impression* sections from the *ith* radiology report projected to a lower dimension, for a batch of N pairs of *Findings* and *Impression* sections their contrastive loss (RSM loss) is defined in Figure 2.2 where τ_1 is a scaling parameter to control the margin and set to $\tau_1 = 0.5$. The final loss of third phase is a weighted combination of RSM loss and MLM loss [4].

$$\mathcal{L}_{\text{RSM}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log \frac{\exp(\mathbf{t}_i^{\text{F}} \cdot \mathbf{t}_i^{\text{I}}/\tau_1)}{\sum_{j=1}^{N} \exp(\mathbf{t}_i^{\text{F}} \cdot \mathbf{t}_j^{\text{I}}/\tau_1)} + \log \frac{\exp(\mathbf{t}_i^{\text{I}} \cdot \mathbf{t}_i^{\text{F}}/\tau_1)}{\sum_{j=1}^{N} \exp(\mathbf{t}_i^{\text{I}} \cdot \mathbf{t}_j^{\text{F}}/\tau_1)}\right)$$

Figure 2.2.: Contrastive Loss introduced in CXR-BERT language model for RSM task [4]

2.5. SciBERT

SciBERT is a BERT [9] based pre-trained language model that is trained on large corpus of scientific data which is a random sample of 1.14M papers from Semantic Scholar [7]. The model has the same architecture, optimization and hyperparameter settings as BERT model [3]. The corpus of the model consists of 18% computer science domain and 82% biomedical domain papers [3]. From the four models that SciBERT introduced, the models that use the corpus of the BERT model are finetuned for NLP downstream tasks such as NER, Text Classification and Dependency Parsing, while the models that use the new scientific corpus are pre-trained from scratch [3]. In addition to outperforming BERT base model on computer science tasks, SciBERT also provides significant improvements on biomedical tasks compared to BioBERT.

The pre-training on a large amount scientific data from scratch in SciBERT enables the model to learn the underlying patterns and structures in scientific data and shows the importance of a domain scientific vocabulary [47].

2.6. Data Augmentation

Data augmentation refers to the transformation techniques that artificially increase the amount of data by adding newly generated synthetic data from already existing data. While data augmentation makes small modifications to existing data, it also increases the variability and diversity of the data effectively [27]. Especially in the cases that the training data is insufficient, data augmentation plays an important role to create additional synthetic data. Even though data augmentation is widely applied in CV for images such as flipping, rotation, and cropping; adaptation of the augmentation techniques to NLP is difficult and under-explored due to discrete structure of natural language [27].

Data augmentation techniques in NLP range from simple noise based methods such as swapping, deletion, insertion to learnable generation based methods such as machine translation. Noise based data augmentation methods in NLP have the focus of adding weak noise to the data that will not change the semantics drastically while creating a slight deviation from the original data. One of the examples for noise based data augmentation methods is swapping on word or sentence level. In some works such as in [58] and [28] , authors randomly choose two words from a number of sentences and swap their positions within sentences as a word level text augmentation strategy [27]. In another study, introduced in [8], authors split the token sequence into several segments, and shuffle the order of tokens in randomly chosen segments while keeping the order of labels unchanged for the segments [27].

Randomly removing each word with a certain probability as performed in [50], [56], and [42], is an example of deletion augmentation strategy in word level. Another example from [50], authors propose to select a random word from a sentence and insert a random synonym of the word into a random position in the sentence which is an approach to perform substitution as an augmentation strategy. One approach to implement substitution is randomly replacing words with their synonyms and hypernyms as a paraphrasing method while keeping the semantics as unchanged as possible. Another way to apply substitution is replacing words with their misspelled versions or applying dropout to a random word and replace the word with a placeholder. This approach improves generalization while reducing the information in the sentence. Replacing words with their k-nearest neighbors that are calculated by cosine similarity of word embeddings is another example of using substitution as data augmentation which was proposed in [48] for Twitter message classification task [27].

With the development of machine translation models, back translation have also become one of the data augmentation strategies in NLP. In back translation, the original data is translated into other languages and then translated back to the original language to create augmented data. Back translation works on sentence level. Therefore, rather than replacing or deleting words, it uses sentences as input and returns an augmented sentence after the augmentation process [27].

Methods such as swapping, deletion, insertion or substitution that act as the state of the art data augmentation strategies on word level can also be applied on sentence level. The approach proposed in [54] performs shuffling, random swapping and random insertion on sentence level for legal document classification task. Authors also perform random deletion of sentences with a certain probability. With these approaches authors aim to increase the scale of their training data in legal domain on sentence level [54]. In CXR-BERT model [4], the sentences of the *Impression* and *Findings* sections of the radiology reports are randomly shuffled before the pre-training in the third phase as an effective text augmentation strategy.

Other than generalization of the data and creating additional data for limited existing data, data augmentation plays an important role in contrastive learning objectives when the task is predicting the input itself in unsupervised setting [11, 27]. In [11], authors take an input sentence and try to predict the sentence in a contrastive learning setting. By applying different standard dropout masks in pre-trained encoder, authors obtain two different embeddings of the same sentence as positive pairs. Emebddings of the other sentences from the same batch are treated as negatives [11]. In SimCLR paper [6], the contrastive learning framework aims to learn the visual representations by maximizing agreement between differently augmented views of the same data example. For this purpose, authors apply random cropping, random color distortions, and random Gaussian blur as data augmentation.

3. Datasets

3.1. The MIMIC-CXR Dataset

The MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0 [30] is a large publicly available dataset of chest radiographs in Digital Imaging and Communications in Medicine (DICOM) format with free-text radiology reports. The dataset aims to support and encourage research in medical imaging and biomedical NLP. It contains 377,110 chest X-ray images from 227,835 radiographic studies. Each radiographic study contains one or more images with their corresponding radiology reports. To protect patient privacy according to the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements, the reports are de-identified by replacing Protected Health Information (PHI) with three consecutive underscores ("___") [30]. Free-text radiology reports in MIMIC-CXR include interpretations of CXR images and summaries of the findings that are prepared by clinicians during routine clinical care and were extracted from the hospital Electronic Health Record (EHR) system. They are semi-structured and contain sections such as *Findings* and *Impression*. While *Findings* section explains the detailed assessment of the corresponding CXR images, *Impression* section acts as a summary of the relevant findings [33]. An example free-text radiology report from MIMIC-CXR is depicted in Figure 3.1.

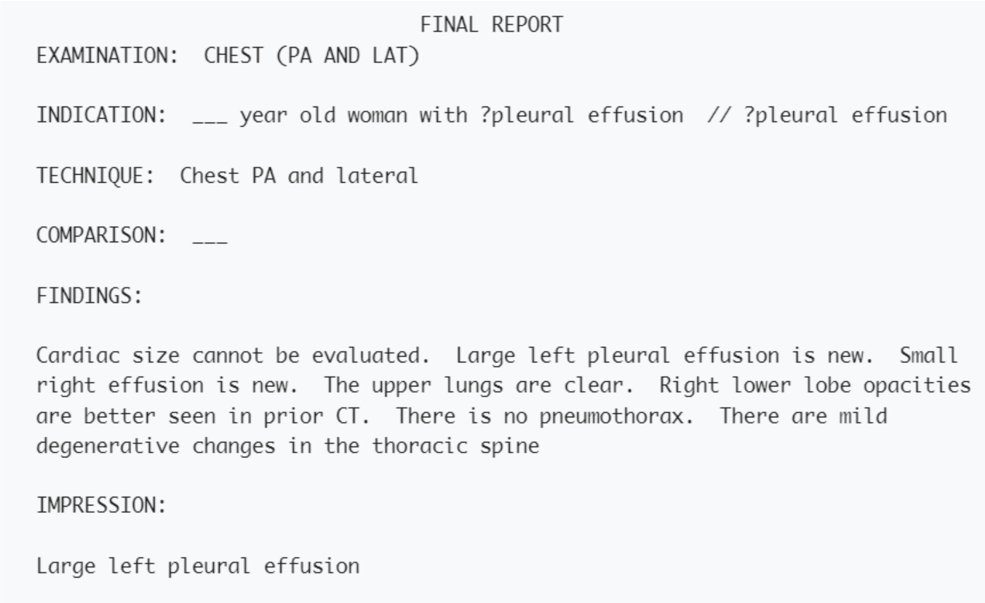
Even though the reports are de-identified, it still contains detailed information regarding the clinical details of patients, therefore it needs to be used with appropriate respect [30].

3.2. Downstream Tasks Datasets

This section includes the explanations of the datasets that have been used in downstream tasks.

3.2.1. MIMIC-CXR Dataset With CheXpert Labels

CheXpert is an open-source rule based labeler that is been used to detect the presence of 14 observations and generate labels from MIMIC-CXR radiology reports [32]. Each mention of the 14 observations is labeled as positive, uncertain, or negative using



```
                                 FINAL REPORT
 EXAMINATION:  CHEST (PA AND LAT)

 INDICATION:  ___ year old woman with ?pleural effusion  // ?pleural effusion

 TECHNIQUE:  Chest PA and lateral

 COMPARISON:  ___

 FINDINGS:

 Cardiac size cannot be evaluated.  Large left pleural effusion is new.  Small
 right effusion is new.  The upper lungs are clear.  Right lower lobe opacities
 are better seen in prior CT.  There is no pneumothorax.  There are mild
 degenerative changes in the thoracic spine

 IMPRESSION:

 Large left pleural effusion
```

Figure 3.1.: An example radiology report from MIMIC-CXR dataset [33]

contextual information. If a positive mention of an observation exists in the associated study, the label for the corresponding observation is assigned as positive and denoted with 1. If an observation is not mentioned in the report, the label is assigned as negative and denoted with 0. When the observation is mentioned with an uncertainty or mentioned with an ambiguous language in the report, the label is assigned as uncertain and denoted with 0. If there is no mention of the observation was made in the report, the label is left blank to denote it is missing [32, 19]. Labels were mostly derived from *Findings* and *Impression* sections of the reports. A total of 227,827 studies are assigned a label by CheXpert labeler. 8 studies could not be labeled due to a lack of a *Findings* or*Impression* section [32].

Final dataset with CheXpert labels consists of 377,110 JPG format chest X-ray images and their corresponding 227,827 free-text radiology reports. They are labeled for the presence of 14 observations as positive, negative, uncertain or missing as a large public dataset for chest radiograph interpretation [32]. 14 observations as labels are presented as Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumonia, Pneumothorax, Pleural Other, Support Devices and No Finding in the corresponding CheXpert paper [19, 32].

3.2.2. MedNLI and RadNLI Datasets

MedNLI [45] is a NLI dataset in biomedical domain annotated by doctors according to the medical history of the patients. Each record in the dataset consists of a premise statement, an hypothesis statement, a gold label, and parsed versions of the statement sentences. The dataset is designed to be used in classifying the premise and hypothesis pairs into one of the three categories as labels: entailment, contradiction, or neutral. Entailment label represents the case that the hypothesis can be inferred from the premise. For contradiction label, the hypothesis cannot be inferred from the premise and the neutral label shows that the inference relation is undetermined [4].

The dataset is in JSON lines format and the source of the premise statements are the clinical notes from MIMIC-III dataset [20]. Each note was segmented into sections and sentences from the "past medical history" section of the clinical notes were randomly sampled for the premise statements. To minimize the risks to patient privacy, clinical notes corresponding to the deceased patients are used for the statements. The dataset includes the training, development/validation and test splits. There are 11232 premise-hypothesis pairs in the training split, 1395 pairs in the development split and 1422 pairs in the test split [45].

RadNLI is a NLI dataset for the radiology domain in which sentence pairs are sampled from the validation section of MIMIC-CXR radiology reports. The sentence pairs were annotated and labeled by one medical expert and one computer science expert. Each pair is annotated twice, swapping its premise and hypothesis sentences, resulting in 960 pairs. The dataset consists of 480 pairs for the validation set and 480 pairs for the test set. The dataset folder contains the development and the test sets in JSON lines format, in which each line contains the id of the NLI pair, the premise sentence, the hypothesis sentence and the NLI label as entailment, contradiction, or neutral [40].

3.2.3. RadGraph Dataset

RadGraph [39] is a dataset of entities and relations of full-text radiology reports from MIMIC-CXR and CheXpert datasets, annotated by board-certified radiologists with an information extraction schema. The train set consists of 425 reports and the development set consists of 75 reports, both from the MIMIC-CXR dataset. The patients associated with reports in the train set do not overlap with any patients associated with reports in the development set. The test set consists of 100 reports, 50 from the MIMIC-CXR dataset and 50 from the CheXpert dataset. The patients associated with reports in the test set do not overlap with any patients associated with the rest of the reports in the dataset. The train, dev and test sets are saved to JSON files and each JSON file holds a dictionary in which keys map to nested dictionaries containing information about the report with a data schema. The data schema defines two broad entity types: $Observation$ and $Anatomy$. The $Anatomy$ entity refers to an anatomical body part that is mentioned in a radiology report, such as a "lung". The $Observation$ entities refer to observations made when referring to the associated radiology image. Observations are associated with visual features, identifiable pathophysiologic processes, or diagnostic disease classifications. For example, an Observation could be "effusion" or more general phrases like "increased". Each Observation has an associated level of uncertainty [39]. The uncertainty levels for each observation entity is defined with three levels: Definitely Present, Uncertain, and Definitely Absent. In total, there are four entities, which are labeled as $ANAT - DP$, $OBS - DP$, $OBS - U$, and $OBS - DA$. Three relations between entities defined in data schema are labeled as $suggestive - of$, $located - at$, and $modify$.

4. Methods

This thesis aims to implement and pre-train a domain specific language model for MIMIC-CXR radiology reports by applying contrastive learning methods on the sections of the reports. For this purpose, the objective of the pre-training follows the same structure with RSM objective from the third phase of the CXR-BERT [4] model.

This chapter gives an overview about the model architecture designed to achieve the training objectives. Afterwards, it dives into each component of the model architecture followed by detailed information for better understanding the main approach behind this work.

4.1. Model Initialization

The language model architecture for pre-training follows the part of the third phase objective of CXR-BERT [4] model. Rather than constructing a custom vocabulary of MIMIC-CXR radiology reports or closely related domain data and pre-training a randomly initialized model to learn the corpora from scratch, introduced approach in this thesis chooses to initialize the weights of the model with a pre-trained model that is from a closely related domain as the baseline model. For this purpose, a custom WordPiece [52] vocabulary of 30k tokens from the radiology reports is constructed and trained from scratch. Constructed vocabulary is compared with corpora of different models that are pre-trained with closely related domain specific data. Results showed that 25% of the tokens from MIMIC-CXR radiology reports vocabulary are common with BioBERT corpora. With SciBERT vocabulary, there are 37% common tokens of MIMIC-CXR radiology reports. By considering vocabulary coverage and the objective of training both cased and uncased models for radiology reports, SciBERT is chosen as the base model to initialize the model weights.

4.2. Data Preprocessing

In this thesis, *Findings* and *Impression* sections of MIMIC-CXR radiology reports are used as training data. As the first step of the data preprocessing, each radiology report is splitted into sections and all section names are normalized. The reports that have empty *Findings* or *Impression* sections are removed from the dataset. The resulting training data set consists of 128032 radiology reports and are stored in a CSV file. Each line of the CSV file represents a radiology report, and it contains a *Findings* section and an *Impression* section in two seperate columns. To split reports into sections, the official documentation published in MIMIC-CXR repository [31] is used.

To prepare the data for augmentation on sentence level as the next step, *Findings* and *Impression* sections of each report are splitted into sentences within sections by using Stanza [36] library for English language. The splitted sentences within sections are stored as lists of sentences for the corresponding sections of each report. The resulting data stored in a CSV file that contains lists of sentences of *Findings* and *Impression* sections in two separate columns, and each line in the CSV file represents a radiology report. This process has been performed before data augmentation on sentence level and the model pre-training.

4.3. Data Augmentation on Sentence Level

This section explains the data augmentation strategies applied on sentence level for the sections of the radiology reports. Each methodology is applied on the fly during the model pre-training. The main purpose of data augmentation methodologies on sentence level is creating an effective text-augmentation strategy in pre-training while introducing permutation invariance on sentence level for the report sections as in CXR-BERT [4]. Data augmentation strategies in this thesis are divided into two main sections as Noise Based Augmentation and Retrieval Based Paraphrasing Augmentation strategies.

4.3.1. Noise Based Augmentation

As one of the noise based augmentation strategies shuffling is applied. The sentences of *Findings* and *Impression* sections of the reports are randomly shuffled within each section as in CXR-BERT [4]. Another noise based augmentation strategy used in this thesis is dropping. For dropping augmentation strategy, random sentences are chosen to be removed from the lists of sentences of a section for each report. This process has been performed for both *Findings* and *Impression* sections. The process also makes sure that both sections still have at least one sentence. Lastly, swapping sentences between sections is used as data augmentation on sentence level. To introduce a slight noise in data distribution and create a better generalization, sentences of *Findings* and *Impression* sections are swapped randomly between sections for each report. This process has been repeated up to the number of sentences in *Impression* section of the each report. Each noise based augmentation strategy is used seperately in the pre-training pipeline of three different uncased and three different cased models.

4.3.2. Retrieval Based Augmentation

This augmentation strategy aims to swap sentences with their nearest neighbors within sections. As the first step, all sentences from *Findings* and *Impression* sections are collected. Results show that there are a total number of 143508 sentences used in *Findings* section with different variations and there are a total number of 61082 sentences from *Impression* sections that are used in different combinations to create the section. The corresponding embeddings of the sentences are collected from SciBERT as the baseline model in the first step and are stored in two files corresponding to *Findings* and *Impression* sections. While the first column in the files stores the sentences, the second column stores the corresponding embeddings of the sentences. During pre-training of the model, a random number of sentences from the list of sentences in each section for each report in training data are swapped with their nearest neighbor from a subset of sentences of the corresponding section. Due to the high demanding time of searching the nearest neighbors for an input sentence, only 5% and 1% of the total number sentences in each section are used to create the subset of sentences as the search database of nearest neighbors for input sentence queries.

To perform an efficient similarity search of sentence embeddings as dense vectors, FAISS [21] library is used. FAISS is a library that has been developed by Facebook AI for efficient similarity search for sets of vectors of any size. In this thesis, FAISS is used to store the sentence embeddings in an index type data structure according to a similarity metric. After the structure is constructed and sentence embeddings are added to the index, the search is performed for the provided query vector in a number of vectors. To fasten the search and decrease the search scope through clustering The Inverted File Index (IVF) is used. In this approach the search space is divided into clusters in which the vectors are assigned according to Euclidean distance to the centroid of the clusters. After the assignment of the vectors to the clusters, the query vector is compared to each of the vectors in the clusters according to Euclidean distance. The number of clusters is set to 100. Each query vector is a sentence embedding from the list of sentences of a section in a report. The number of nearest neighbors to retrieve is set to 2 to create a better generalization in the augmentation process. After the search in the index data structure is completed and the nearest neighbors of the query vector are found, the corresponding sentence to the query vector is replaced randomly with one of the sentences corresponding to nearest neighbor vectors. The augmentation is performed for both *Findings* and *Impression* sections. Using FAISS library provides an efficient similarity search by speeding up the search time and fastening the augmentation process.

The augmentation has been applied on the fly during pre-training and repeated for each epoch. In this way, updated model weights in each epoch are used for the sentences of each section and a dynamic data augmentation is performed.

4.4. Model Architecture and Pre-Training Objective

The proposed model architecture has the same structure with the contrastive learning part of third phase of the CXR-BERT model depicted in Figure 2.1. Different than CXR-BERT, the proposed language model is trained only with RSM objective and the model weights are initialized with the SciBERT weights rather than pre-training from scratch. The text encoder of the language model is based on the BERT [9] base size architecture. On top of the text encoder, there is a projection layer P_{txt} which is a two-layer perceptron to project the 768-dimensional feature vector of $[CLS]$ tokens to a 128-dimensional representations as in CXR-BERT [4].

The training objective that aims to match *Findings* and *Impression* sections of the same report uses contrastive loss that favours *Findings* and *Impression* pairs from the same report over pairs from different reports. Following a similar pipeline with CXR-BERT and assuming there are D set of pairs of *Findings* and *Impression* radiology report sections, let w^F denote a vector of T (sub-)word tokens of the *Findings* section of a report x_{txt} and w^I denote a vector of T (sub-)word tokens of the *Impression* section of the report x_{txt} after the data augmentation and tokenization. Further, let $(\tilde{t}^F, \tilde{t}^I)$ denote a pair of $[CLS]$ tokens corresponding to the *Findings* and *Impression* sections of the same report, and let (t^F, t^I) denote the pair projected to a lower dimension via the projection layer P_{txt}. For a batch of N projected pairs, the contrastive loss in the pre-training as RSM loss is defined in CXR-BERT [4] with the formula in Figure 2.2 and used in the pre-training for this thesis as cost function. In the equation, $\tau_1 > 0$ is a scaling parameter to control the margin that is set to $\tau_1 = 0.5$ as in CXR-BERT.

The pipeline of the pre-training is depicted in Figure 4.1.

4.5. Implementation Details

Following a similar set of hyperparameters to CXR-BERT, the proposed language model uses the AdamW optimiser with a batch size of 128 sequences and a linear learning rate schedule over 50 epochs with a %3 warm up period. The base learning rate is set to 2e-5. The projection layer P_{txt} projects the $[CLS]$ token that is a 768-dimensional feature vector $\tilde{t}$ to a 128-dimensional report representation t. The maximum length of the tokenizer is set to 256. Models with noise based augmentation strategies are

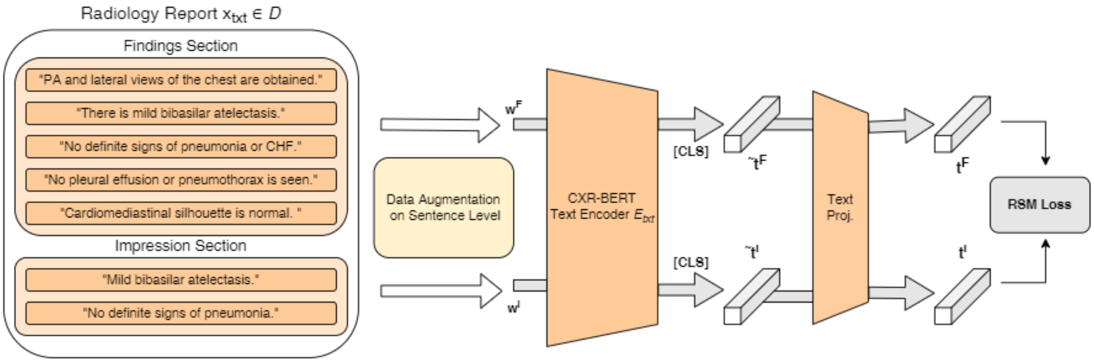

Radiology Report $x_{txt} \in D$
Findings Section
"PA and lateral views of the chest are obtained."
"There is mild bibasilar atelectasis."
"No definite signs of pneumonia or CHF."
"No pleural effusion or pneumothorax is seen."
"Cardiomediastinal silhouette is normal. "
Impression Section
"Mild bibasilar atelectasis."
"No definite signs of pneumonia."
Data Augmentation on Sentence Level
w^F
w^I
CXR-BERT Text Encoder E_{txt}
[CLS]
[CLS]
$\tilde{t}^F$
$\tilde{t}^I$
Text Proj.
t^F
t^I
RSM Loss

Figure 4.1.: The proposed language model pre-training pipeline leverages the effective data augmentation strategies on sentence level and uses RSM loss as the cost function.

trained for 12 hours and models with retrieval based augmentation strategy are trained for 24 hours. All models are trained on a single GPU. Results of downstream tasks for uncased models are collected before the pre-training of the cased models. After observing that the percentage of searched data for nearest neighbors of each section do not have a significant effect on the downstream tasks results, only 1% of section sentences are used for sentence level nearest neighbor data augmentation strategy in cased model pre-training.

5. Downstream Tasks

The main purpose of downstream tasks is evaluating the performance of pre-trained language models in this thesis and comparing with baseline models by finetuning the models for each task. Therefore, each task is simplified for an easy evaluation process. The finetuning pipelines for each task utilize *AutoModel* module from *HuggingFace* library [51]. Each task is performed for cased and uncased models.

5.1. Multi Label Classification of CheXpert Labels

The multi label classification task aims to classify MIMIC-CXR radiology reports according to the presence of 14 different observations in reports that represent chest and lung diseases as labels. For this task, MIMIC-CXR reports with CheXpert labels described in Section 3.2.1 are used. Each label can have four classes as blank, positive, negative, and uncertain.

To simplify the comparison between pre-trained language models in this thesis and existing baseline models, and handle the uncertain and blank classes, a multi label binary classification downstream task is formalized. Following the two common uncertainty handling strategies, $U - Ones$ and $U - Zeros$ approaches [18], blank labels are mapped to negative class (labeled as 0), and uncertain labels are mapped to positive class (labeled as 1) for this task. After the mapping process of the CheXpert labels, a subset of 5000 radiology reports from the train data of the MIMIC-CXR radiology reports with CheXpert labels are chosen randomly as the train data for the finetuning. All data samples in the validation and test sets of the reports with CheXpert labels are used as validation and test sets of the task. The validation dataset consists of 3269 radiology reports with CheXpert labels while test set consists of 1808 samples.

For the model architecture, $AutoModelForSequenceClassification$ class from the $Huggingface\ Transformers$ library is used [51]. The general model architecture for this task includes a classification head on top of the base model encoder. There are 14 classification heads, that are being trained for binary classification of each 14 observations. The models are finetuned for 4 epochs with a learning rate of $3e - 5$ and the batch size of 32. Since the task is a multilabel binary classification task, *Sigmoid* function is applied to the predictions of each observation. As evaluation metrics, F1 scores and

ROC-AUC scores are calculated. The evaluation results of the task are discussed in Section 6: Evaluation and Results in detail.

To analyse how the models pre-trained in this thesis perform compared to the baseline models in different amounts of train data setting and to create comparable baselines for zeroshot classification task of the labels, random subsets of 500 and 50 samples from the train data of MIMIC-CXR radiology reports with CheXpert labels are collected. The subsets correspond to 10% and 1% of the full train data in finetuning. The results for subsets are presented in Section 6: Evaluation and Results and in Appendix A.

5.2. Zero Shot Classification of CheXpert Labels

Zero Shot Classification is the task of classifying samples with a label that was not seen during pre-training [29]. After observing that language models pre-trained in this thesis achieved better performance than baseline models in multi label classification task, test set of the MIMIC-CXR radiology reports with CheXpert labels is used to evaluate the zero-shot classification performance of the language models for multi label classification of the labels. As the prompts of the Zero Shot Classification task, the template from CXR-BERT that consists $"findings\ suggesting\ [observation]"$ and $"no\ evidence\ of\ [observation]"$ sentence tuples is used to represent positive and negative classes. For each one of the 14 observations $[observation]$ word is replaced with the corresponding observation and positive and negative prompts are generated. In next step, the positive and negative prompt embeddings are retrieved for each observation from the pre-trained language model and the cosine similarity between radiology report embeddings and the embeddings of the positive and negative prompts is calculated. For each observation Sigmoid function is used to turn the cosine similarities into probability scores, and the presence of the observation in the radiology reports is classified according to the thresholds defined for each observation. Choice of the thresholds and results of the task are discussed in Section 6: Evaluation and Results.

5.3. Natural Language Inference (NLI) for RadNLI Dataset

NLI task aims to determine whether a provided hypothesis can be inferred from a provided premise [43]. To evaluate the performance of the pre-trained models in this thesis compared to baseline models in biomedical domain, RadNLI [40] and MedNLI [45] datasets are used in this task. As explained in Datasets 3.2.2 section, RadNLI dataset consists of labelled hypothesis and premise sentence pairs sourced from MIMIC-CXR radiology reports. Each pair is categorized with one of the *entailment*, *contradiction* and *neutral* labels. RadNLI dataset provides 480 sentence pairs for each validation and test sets, but there is no official train set. Therefore, MedNLI dataset that has 11k labelled hypothesis and premise sentence pairs is used as train set of the downstream task.

The general model architecture for this task includes a classification head on top of the base model encoder to classify each pair with one of the 3 labels. Since the task is a binary classification task with 3 labels, Softmax function is applied to the predictions. For the model architecture, $AutoModelForSequenceClassification$ class from $Huggingface\ Transformers$ library [51] is used.

Following the CXR-BERT hyperparameter settings for the task, the language models are finetuned up to 20 epochs and early stopping is used by monitoring accuracy scores on the RadNLI development set. The stopping patience is set to 5, learning rate is set to $1e-5$ and batch size is 16. As evaluation metrics, accuracy, F1 score, precision and recall values are calculated. The evaluation results of the task are discussed in Section 6: Evaluation and Results in detail.

5.4. Named Entity Recognition (NER) for RadGraph Dataset

Named Entity Recognition (NER) for RadGraph Dataset aims to identify and classify named entities provided in RadGraph dataset to retrieve clinically relevant information within the radiology reports. As described in section 3.2.3, RadGraph dataset includes the entities and relations from MIMIC-CXR radiology reports. In this task, train and development sets of from RadGraph dataset is used as train and validation data for finetuning. For the test set, only MIMIC-CXR radiology reports with entities from RadGraph dataset is used. In tagging process of the chunk of tokens with corresponding entities, the Inside-Outside-Beginning (IOB2) tagging format [41] is used which is a common tagging format for tagging tokens in chunks. As the first step, entities for each corresponding token from RadGraph dataset are tagged with O tag and $I-$ or $B-$ prefixes. Afterwards, tokens are aligned with their corresponding entities by using start and end indexes of the chunks. The O tag indicates that the token does not belong to any entity and therefore does not have an assigned entity type. $B-$ prefix indicates the beginning of an entity for the corresponding token in a chunk of tokens. $I-$ prefix shows that the corresponding token for the entity is not the beginning token but is inside the current chunk of tokens and has the same entity with the current chunk of tokens. There are 8 labels after the mapping process with the beginning and intermediate tags for each entity and O tag for no entity type. Constructed labels are namely *B-ANAT-DP, I-ANAT-DP, B-OBS-DP, I-OBS-DP, B-OBS-DA, I-OBS-DA, B-OBS-U, I-OBS-U*.

After tagging and alignment process of the entities with tokens, the models are initialized for finetuning. Different than other downstream tasks that works on sequence level, NER task is a classification task on token level. Therefore, $AutoModelForTokenClassification$ class from $Huggingface\ Transformers$ library [51] is used for the model architecture. The learning rate is set to $3e-5$ and models are finetuned for 3 epochs. After observing that the task becomes too easy for the models when the whole train data of RadGraph dataset is used, only 25% of the train data is chosen as the train subset for the task. The samples in the subset are selected randomly from the RadGraph train data. As evaluation metrics accuracy and F1 scores are collected. The results of the task and evaluation strategy of entities are discussed in Section 6: Evaluation and Results in detail.

6. Evaluation and Results

This chapter provides information about collected results from experiments and the evaluation strategies that are chosen for each downstream task.

6.1. Multi Label Classification of CheXpert Labels

The mapping of the uncertain and blank observations to positive and negative classes shows that the final dataset for finetuning is an imbalanced dataset in which for each observation there are less mention of the observation (positive class) than no mention of it (negative class). After observing that using the default threshold as 0.5 for the classification task does not provide comparable results especially for the 10% and 1% subsets as train data for finetuning, an evaluation strategy that focuses on threshold tuning for each observation label independently is introduced. In this strategy precision, recall and threshold from the precision and recall curve for each observation are calculated in each epoch for the validation set and stored. At the end of the finetuning, corresponding thresholds for the precision and recall values that give the best F1 score are selected as the best thresholds for each observation independently. This strategy aims to create a comparable baseline of results for different models while keeping the original ROC-AUC scores unchanged. In Table 6.1, F1 scores and ROC-AUC scores of uncased models are presented for the case that 100% of the finetuning train data is used. Same experiment is repeated for cased models and results are presented in Table 6.2.

Results show that uncased models that are pre-trained in this thesis with different data augmentation strategies improve the initial SciBERT model performance. While the models with swapping and dropping sentences augmentation strategies achieve 8% improvement in F1 score as the performance metric, the model with the nearest neighbor augmentation strategy that uses 5% of all sentences in neighbor search improves the initial SciBERT model performance 14% for the case that 100% of the finetuning train data is used. Additionally, uncased models pre-trained in this thesis achieve better results than other baseline models in terms of F1 scores for multi label binary classification of CheXpert labels when different amounts of train data is used during finetuning. For uncased models, F1 score of the initial model SciBERT is improved between $4 - 16\%$ for 10% train data subset and 18% improvement in F1 score is observed in experiments with 1% train data subset. In Figure 6.1, performance of the uncased models are depicted with three different train data percentage.

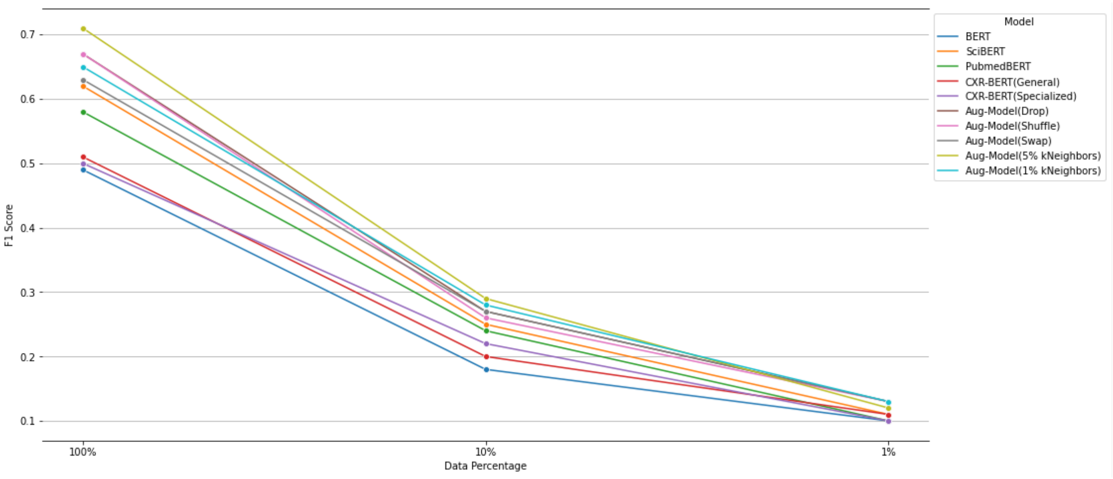


Figure 6.1.: F1 Scores for multi label classification of CheXpert labels for uncased models with different amounts of train data sets in finetuning. Aug-Model correspond to the pre-trained language models in this thesis.

For the cased models pre-trained in this thesis, the experiment with 100% of the finetuning train data gives the same results with the initial model, SciBERT. Similar to performance improvement of uncased models, the experiments with 10% subset of the finetuning train data for the cased models show that cased models pre-trained in this thesis with shuffling and swapping sentences augmentation strategies achieve 9% improvement in F1 score as performance metric. Results are presented in Table: A.1 for uncased models and Table A.3 for cased models. In Figure 6.2, performance of the cased models are depicted with three different train data percentage in finetuning.

Experiments with the 1% train data subset for finetuning result with unsatisfactory performance for both uncased and cased models. Even though F1 scores of uncased and cased models pre-trained in this thesis are better than other baseline models, ROC_AUC scores for both uncased and cased models show that models predict close to random with low ROC_AUC scores. Results for uncased and cased models that use 1% train data subset in finetuning are presented in Table: A.2 for uncased models and in Table: A.4 for cased models in Appendix. Results collected in this task are also used in zeroshot classification of the CheXpert labels described in the following section.

Model	F1 Macro	F1 Weighted	ROC-AUC Macro	ROC-AUC Weighted
BERT	0.49	0.75	0.9	0.95
SciBERT	0.62	0.87	0.94	0.96
PubmedBERT	0.58	0.84	0.93	0.96
CXR-BERT(General)	0.51	0.67	0.85	0.91
CXR-BERT(Specialized)	0.5	0.73	0.83	0.92
Aug-Model(Drop)	0.67	**0.88**	**0.96**	**0.98**
Aug-Model(Shuffle)	0.67	0.87	0.94	0.97
Aug-Model(Swap)	0.63	0.87	0.94	0.97
Aug-Model(5%-kNeighbors)	**0.71**	**0.88**	0.95	0.97
Aug-Model(1%-kNeighbors)	0.65	0.87	0.94	**0.98**

Table 6.1.: Test set results multi label classification of CheXpert labels for uncased models with 100% of train data in finetuning. Aug-Model models correspond to the pre-trained models in this thesis with sentence level data augmentation strategies. Bold and underlined indicates the best result overall.

Model	F1 Macro	F1 Weighted	ROC-AUC Macro	ROC-AUC Weighted
BERT	0.72	0.87	0.92	0.97
SciBERT	**0.84**	**0.94**	**0.96**	**0.98**
BioBERT	0.75	0.89	0.94	**0.98**
Bio-ClinicalBERT	0.78	0.91	0.95	**0.98**
Aug-Model(Drop)	0.83	0.93	**0.96**	**0.98**
Aug-Model(Shuffle)	**0.84**	**0.94**	**0.96**	**0.98**
Aug-Model(Swap)	**0.84**	**0.94**	**0.96**	**0.98**
Aug-Model(5%-kNeighbors)	**0.84**	**0.94**	**0.96**	**0.98**
Aug-Model(1%-kNeighbors)	0.82	0.93	**0.96**	**0.98**

Table 6.2.: Test set results multi label classification of CheXpert labels for cased models with 100% of train data in finetuning. Aug-Model models correspond to the pre-trained models in this thesis with sentence level data augmentation strategies. Bold and underlined indicates the best result overall.

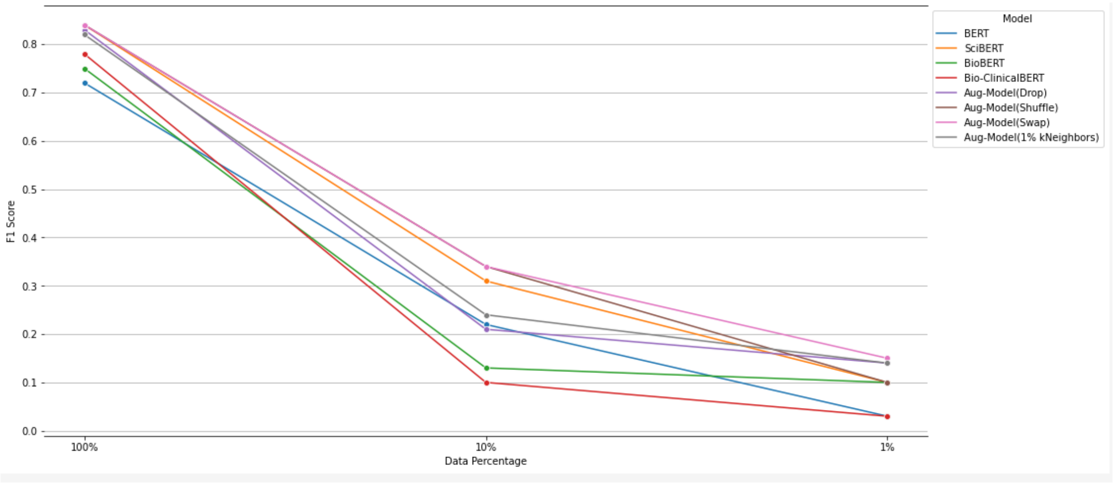


Figure 6.2.: F1 Scores for multi label classification of CheXpert labels for cased models with different amounts of train data sets in finetuning. Aug-Model correspond to the pre-trained language models in this thesis.

6.2. Zero Shot Classification of CheXpert Labels

For the evaluation of zero-shot classification of CheXpert labels task, the best thresholds for each observation that have been found during threshold tuning in multi label classification task are used. After the positive and negative classes are assigned according to the corresponding thresholds for each observation, F1 scores and ROC-AUC scores are calculated. Results for the baseline models with 10% and 1% train subsets in finetuning from multi label classification of CheXpert labels are used to compare zeroshot performance of the pre-trained models in this thesis.

The results show that the uncased and cased language models pre-trained in this thesis have low ROC-AUC scores that are close to random predictions and do not perform well for classifying MIMIC-CXR radiology reports according to the presence of 14 observations without being finetuned on the the radiology reports with CheXpert labels. Experiment results of uncased and cased baseline models such as SciBERT and BioBERT with 10% and 1% subsets of train data in finetuning also prove that when the amount of train data decreases in finetuning, predictions become closer to random for the multi label classification task. A possible reason for the performance decrease is the imbalanced data structure of CheXpert labels of MIMIC-CXR radiology reports after mapping process of uncertain and blank values for observations. While during finetuning, models have the opportunity to learn the distribution of the labelled data in imbalanced setting, zero shot classification task does not perform any finetuning and therefore models fail to classify reports with correct observation labels. Similar results are seen for both uncased and cased models.

Another possible reason of close to random results for zeroshot classification is the provided positive and negative prompts. The retrieved embeddings of the prompts from the language models are close to each other in embedding space, therefore assigned probability of the prompts are close to 0.5 for each observation which requires further research to understand how different prompts effect the zeroshot classification performance in imbalanced data setting for biomedical domain.

6.3. Natural Language Inference (NLI) for RadNLI Dataset

As evaluation strategy accuracy, F1 score, precision and recall values for the NLI downstream task is collected for the models.

Model	Accuracy	F1 Macro	Precision	Recall
BERT	0.38	0.55	0.54	0.58
SciBERT	0.55	0.63	0.61	0.67
PubmedBERT	0.55	0.63	0.60	0.67
CXR-BERT(General)	0.60	**0.67**	**0.64**	**0.72**
CXR-BERT(Specialized)	**0.64**	0.64	0.62	0.68
Aug-Model(Drop)	0.57	0.64	0.61	0.68
Aug-Model(Shuffle)	0.51	0.61	0.59	0.64
Aug-Model(Swap)	0.52	0.61	0.59	0.65
Aug-Model(5%-kNeighbors)	0.55	0.59	0.56	0.63
Aug-Model(1%-kNeighbors)	0.52	0.62	0.60	0.66

Table 6.3.: Test set results from natural language inference of RadNLI dataset for uncased models. Aug-Model models correspond to the pre-trained language models in this thesis.

Uncased models pretrained in this thesis with different augmentation strategies perform similar to the initial model SciBERT and PubmedBERT in terms of F1 score and accuracy. Different than other pre-trained models in this thesis, the pre-trained model with dropping sentences augmentation strategy performs 3% improvement on initial SciBERT accuracy. Additionally, CXR-BERT models achieve the best performance and the results collected in this thesis for this task are aligned with the results provided in corresponding paper [4]. One possible reason for similar performance results of the uncased models pre-trained in this thesis with the initial model SciBERT rather

Model	Accuracy	F1 Macro	Precision	Recall
BERT	0.42	0.6	0.6	0.61
SciBERT	0.48	0.6	0.58	0.63
BioBERT	0.47	**0.63**	**0.63**	**0.65**
Bio-ClinicalBERT	0.5	0.57	0.54	0.62
Aug-Model(Drop)	0.51	0.61	0.6	0.64
Aug-Model(Shuffle)	**0.53**	0.6	0.57	0.64
Aug-Model(Swap)	0.51	0.6	0.57	0.63
Aug-Model(1%-kNeighbors)	0.51	0.6	0.58	0.63

Table 6.4.: Test set results from natural language inference of RadNLI dataset for cased models. Aug-Model models correspond to the pre-trained language models in this thesis.

than better performance is catastrophic forgetting which is a significant problem in continuous pre-training settings. Catastrophic forgetting is the inability of a model to retain previous information in the presence of the new information [22]. In other words, when a model is trained with new information that interferes with previously learnt knowledge, it can lead to a performance decrease or, in the worst case, can cause the old knowledge being completely overwritten by the new one [35]. After being trained on RSM task, SciBERT weights that have been used as the initial weights might be overridden by the pre-training task during continous pre-training. As a possible result from the override, models pre-trained in this thesis lose their generalization ability on the task and they perform slightly degraded compared to SciBERT. Considering the fact that models pre-trained in this thesis are trained only for 12 hours and SciBERT is already a high performing model, the improved effects of the continuous pre-training and later finetuning for this task are not seen in uncased pre-trained models. Results are presented in Table: 6.3 for uncased models.

Another possible reason is the difference between MedNLI dataset used as finetuning train data for the task and MIMIC-CXR reports used as the pre-training data. While other tasks use subsets that have been constructed from MIMIC-CXR radiology reports with labels, hence work with a semi-supervised approach, the NLI task follows an approach that is more close to transfer learning due to the use of MedNLI dataset. As explained in Section 3.2.2, MedNLI dataset is constructed from MIMIC-III dataset. The uncased and cased models introduced in this thesis are pre-trained only on MIMIC-CXR radiology reports. During finetuning, MEDNLI dataset introduces a new data distribution that is different than the pre-training data. As a possible result of the difference, uncased pre-trained models struggle to represent and understand the new data with previously learned representations during pre-training. Therefore they perform slightly poorer than initial model in this task. Additionally, choice of hyperparameters can also cause performance degradation. To be able to replicate the same results and compare the uncased pre-trained models in this thesis with CXR-BERT, same hyperparameters for the NLI task from CXR-BERT is used in finetuning. In the case that provided hyperparameters are not well-suited for the pre-trained models in this thesis, models perform poorer.

For cased models pre-trained in this thesis, the difference in data distribution betweeen MIMIC-CXR radiology reports and MedNLI dataset and catastrophic forgetting do not cause a performance decrease. Results show that cased models pre-trained in this thesis achieve 6% improvement in accuracy as the performance metric of the task compared to the initial model SciBERT. While results of the other tasks for the models pre-trained in this thesis with different augmentation strategies are close to each other, the pre-trained model with shuffling data augmentation strategy performs better than other pre-trained models and achieves 10% improvement in accuracy. Results are presented in Table 6.4 for cased models.

6.4. Named Entity Recognition (NER) for RadGraph Dataset

The evaluation strategy for the task considers entity level evaluation rather than collecting results on token level. Therefore, measured results for precision, recall and F1 score are collected at entity level by following the evaluation strategy from SemEval 2013-9.1 task [44]. Used evaluation metrics consider six different scenarios for the correct entity type classification of the tokens and correct boundary assignment of the classified entity types to the tokens. For this purpose, *nervaluate* module [2] is used which is a python module for evaluating NER tasks according to the defined strategies in [44]. Scenarios include the cases such as boundary and entity type match in which both are correctly predicted. Other scenarios are incorrect entity prediction, missed entities, wrong entity type assignment with correct boundary prediction for the token chunks, wrong boundary prediction with correct entity type and wrong entity prediction with wrong boundary assignment. In the evaluation process, there are five different error types as metrics that are used to consider different categories of errors. They are defined as *Correct, Incorrect, Partial, Missing* and *Spurius*. Each metric is measured in four different ways as *Strict, Exact, Partial* and *Type*. Detailed explanations about the evaluation schema and metric types are explained in tables A.7 and A.8 in Appendix. As explained in 5.4, experiments that use the whole train data in finetuning show that the results from each model are similar due to the simplicity of the task. To retrieve comparable results and evaluate the performance of the models pre-trained in this thesis only 25% of the train data of RadGraph dataset is used in this task. Samples in the subset are chosen randomly to form the train set for finetuning.

Results in Figure 6.3 present averaged F1 score from the four evaluation schemas of each entity for each uncased language model. Results show that uncased pre-trained models in this thesis perform similar to SciBERT and PubmedBERT for $ANAT - DP$ and $OBS - U$ entity types while performing poorer for $OBS - DA$ and $OBS - DP$ entity types. Another observation from the experiment results is the performance degradation of specialised CXR-BERT model compared to general CXR-BERT model. Specialised CXR-BERT model is continually pre-trained from general CXR-BERT to further specialize it in chest X-ray domain and to use the final model in multi-modal contrastive learning approach for VLP [4]. Experiment observations show that as a possible result of continuous pretraining, the CXR-BERT model loses its generalization ability and perform poorly for the task. Similar observation can be done for the pre-trained models in this thesis. As a possible reason for the performance decrease for uncased and cased models that are continuously pre-trained from SciBERT, models lose their generalization ability by being trained only on radiology reports and hence have poorer performance compared to SciBERT, which can also be interpreted as catastrophic forgetting similar to NLI task observations.

Additionally, the finetuning train data distribution shows that there are more $ANAT - DP$ entities exist in data and the amount of $OBS - U$ entities are less than other entities. For the case that there is enough data, which is the case for $ANAT - DP$ entity in RadGraph dataset, the task becomes easy for all uncased and cased models and the good performance of the models can not be differentiated and compared. Similar result is also observed for $OBS - U$ entity in the dataset. When there is not enough data and the models do not have the opportunity to learn the data distribution, they perform poorly and it is not possible to compare the performance of the models in terms of F1 score. Results in Figure 6.4 present averaged F1 score from the four evaluation schemas of each entity for each cased model. From the evaluation with F1 scores, it can be observed that results of $OBS - DA$ and $OBS - DP$ entities are comparable for both uncased and cased models. Accuracy of the models are presented in Table: A.5 for uncased models and in Table: A.6 for cased models in Appendix.

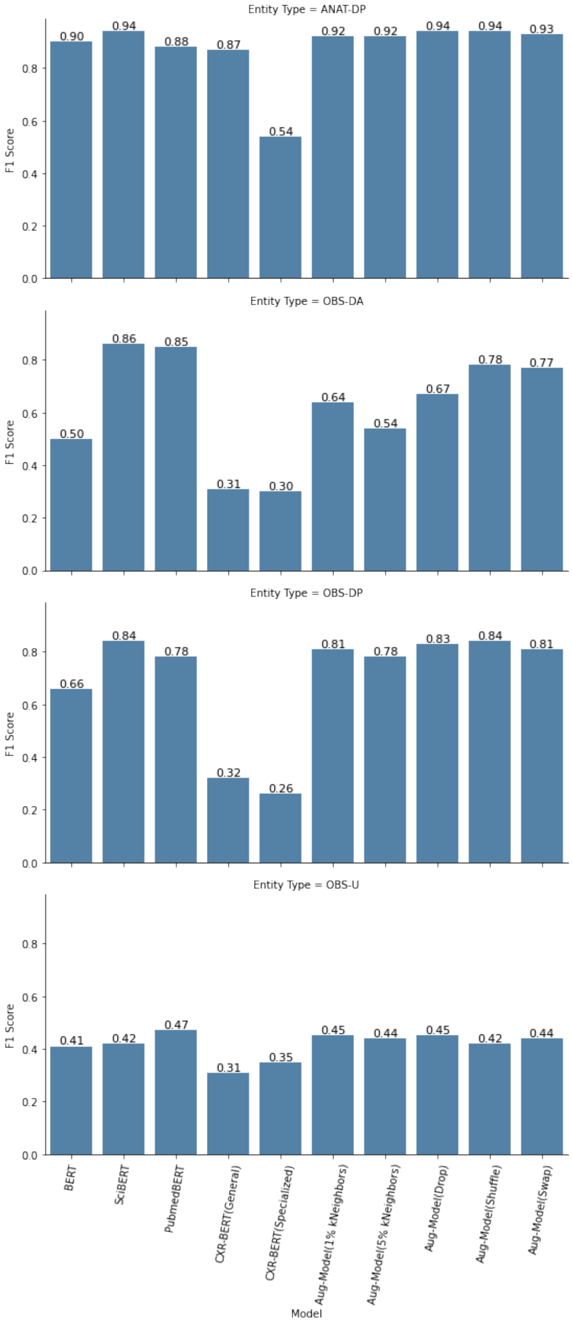


Figure 6.3.: F1 scores for named entity recognition (NER) of RadGraph dataset with each entity type for uncased models. Aug-Model correspond to the pre-trained language models in this thesis.

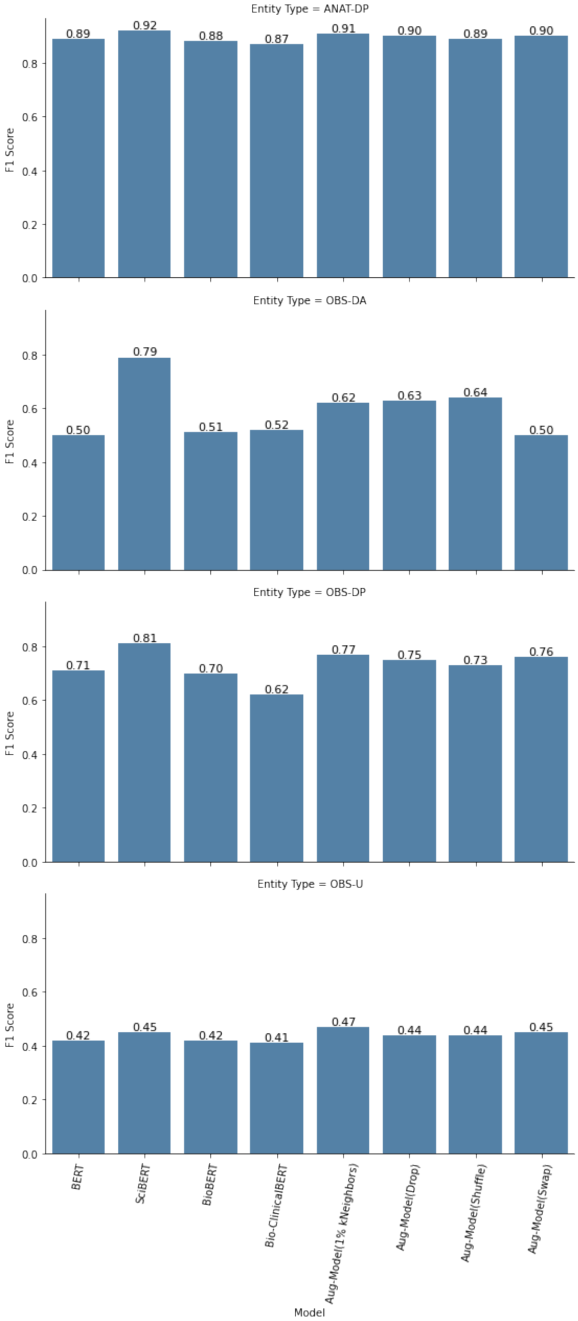


Figure 6.4.: F1 scores for named entity recognition (NER) of RadGraph dataset with each entity type for cased models. Aug-Model correspond to the pre-trained language models in this thesis.

7. Conclusion and Future Work

7.1. Conclusion

In this thesis, a model pre-training pipeline for MIMIC-CXR radiology reports is implemented with a contrastive learning objective. The pre-training process not only leverages the semi-structured nature of the radiology reports, but also performs data augmentation on sentence level. To fulfill the objectives described in Section 1.2, the model architecture and the contrastive learning objective of the pre-trained models in this thesis follow CXR-BERT model structure. Different than the CXR-BERT training pipeline that includes pre-training the model from scratch with biomedical domain data and specializing on radiology reports in further training phases, the models pretrained in this thesis use SciBERT as the initial model and further train SciBERT on radiology reports. By initializing the model weights with SciBERT weights rather than random weights, the pre-trained models not only adapt to biomedical domain knowledge easier due to similar domain of SciBERT, but also reduce environmental costs of training process. After the model is initialized and reports are seperated into sections, data augmentation on sentence level is applied as explained in Section 4.3. In addition to shuffling sentences within sections as performed in CXR-BERT, the pretrained models in this thesis utilize dropping sentences within sections, swapping sentences between sections and replacing sentences with their nearest neighbors within sections as data augmentation strategies on sentence level. The data augmentation strategies applied in the pre-training process are inspired by word level data augmentation strategies. After the data augmentation, the models are pre-trained for RSM objective with contrastive learning approach. Lastly, each pre-trained model is finetuned with four downstream tasks.

The downstream tasks aim to evaluate the performance of the pre-trained models in this thesis and compare the results with state of the art baseline models in biomedical NLP in terms of several evaluation metrics. As presented in Results section 6, the uncased pre-trained models in this thesis achieve performance improvements compared to baseline models for multi label classification of CheXpert labels task in different train data settings during finetuning. On the other hand, cased models pre-trained in this thesis perform similar to initial model SciBERT. Zeroshot classification results show that both uncased and cased pre-trained models in this thesis are not capable enough to capture the differences between mention and no mention of the CheXpert labels as observations. Performance of the baseline models such as SciBERT in this task also show that the predictions are close to random and the task itself requires further research in terms of prompt engineering in biomedical domain and zeroshot performance improvements for imbalanced data. NLI task results show that the uncased pre-trained models in this thesis perform similar to initial model, SciBERT. Possible reasons for the similar performance include catastrophic forgetting, losing generalization ability with further training and choosing ill-suited hyperparameters. For cased models in the same task, pre-trained models in this thesis improve the performance of the base model. Similar to multi label classification task, results for NER provides an overview about the performance change of the models in different amounts of data setting. While models pre-trained in this thesis perform similar to each other but poorer than the initial model, SciBERT, collected results for SciBERT show that it is a stable and high-performing model in biomedical domain.

The application of contrastive learning in semi structured radiology reports and data augmentation on sentence level present a novel pre-training pipeline for biomedical domain data. Collected results from the downstream tasks not only provide insights about the tasks, but also prove the importance of further research needed in biomedical NLP.

7.2. Future Work

Although the pre-training objectives and results from the downstream tasks provide promising results for biomedical NLP, future work is expected to address the issues mentioned in 6: Evaluation and Results section for the downstream tasks through the findings and experiments of this thesis. A possible direction is exploring prompt engineering for better usage of language models in zero shot classification. Specialised CXR-BERT model after the joint model training provides improved results for understanding the shared latent information between radiology reports and CXR images [4]. Even though most experiments for joint image-text training result with improved and comparable results and do not require extensive human expert knowledge in biomedical domain as presented in CXR-BERT, experiments that focus solely on textual data show that downstream task performance for biomedical NLP can heavily depend on the choice of text prompts. Therefore, constructing good and effective text prompts still create a demand for expert domain knowledge which is costly and time-consuming and therefore requires further research [49].

Another possible direction to take into consideration for further exploration according to the experiments conducted in this thesis is the effectiveness of continous pre-training for radiology reports on a stable and high performing model. By continually pre-training SciBERT on unlabelled radiology reports with contrastive learning approach, domain adaptive pre-training [15] is performed. To analyse the domain similarity before pre-training, vocabularies of the domain specific language models and corpora of radiology reports are compared. Due to the higher domain similarity of radiology reports with SciBERT compared to other domain specific models, the potential of domain adaptive pre-training is not leveraged and improved enough compared to the language models that have less domain similarity with radiology reports. This results with a possible research direction for utilizing language models in task adaptive pre-training environment for biomedical NLP. Task adaptive pre-training refers to pre-training on the unlabeled train data for a given task [15]. By observing its effectiveness in different works [17], and considering the fact that it focuses more on task specific data and use less computational resources, approaches on contrastive learning objectives in a task adaptive environment and during fine-tuning can be considered as further exploration areas in biomedical NLP [14].

Last but not least, data augmentation can be considered as one of the possible directions for further research. Data augmentation approaches on sentence level in this thesis aim to contribute to the exploration of augmentation strategies in NLP while introducing permutation invariance for pre-training. Most of the data augmentation approaches focus on completion of mask tokens hence work on token level, sentence level data augmentation strategies and generation of more diverse and high-quality data with less cost still require further research especially in biomedical NLP in which they are under-explored.

List of Figures

List of Tables

A. Appendix

Model	F1 Macro	F1 Weighted	ROC-AUC Macro	ROC-AUC Weighted
BERT	0.18	0.23	0.55	0.57
SciBERT	0.25	0.37	0.66	0.67
PubmedBERT	0.24	0.32	0.57	0.56
CXR-BERT(General)	0.20	0.26	0.56	0.58
CXR-BERT(Specialized)	0.22	0.28	0.65	0.68
Aug-Model(Drop)	0.27	**0.45**	0.68	0.70
Aug-Model(Shuffle)	0.26	0.42	0.68	0.71
Aug-Model(Swap)	0.27	0.44	0.68	0.70
Aug-Model(5%-kNeighbors)	**0.29**	0.44	**0.69**	**0.72**
Aug-Model(1%-kNeighbors)	0.28	0.44	0.68	0.71

Table A.1.: Test set results multi label classification of CheXpert labels in 10% train data setting for finetuning of uncased models. Aug-Model correspond to the models pre-trained in this thesis with sentence level data augmentation strategies. Bold and underlined indicates the best result overall.

Model	F1 Macro	F1 Weighted	ROC-AUC Macro	ROC-AUC Weighted
BERT	0.1	0.1	0.5	0.5
SciBERT	0.11	0.16	**0.54**	0.52
PubmedBERT	0.1	0.11	0.51	0.5
CXR-BERT(General)	0.11	0.1	0.5	0.5
CXR-BERT(Specialized)	0.1	0.12	0.53	0.52
Aug-Model(Drop)	**0.13**	**0.21**	**0.54**	**0.57**
Aug-Model(Shuffle)	**0.13**	**0.21**	**0.54**	0.56
Aug-Model(Swap)	**0.13**	0.2	**0.54**	0.56
Aug-Model(5%-kNeighbors)	0.12	**0.21**	**0.54**	0.56
Aug-Model(1%-kNeighbors)	**0.13**	**0.21**	**0.54**	**0.57**

Table A.2.: Test set results multi label classification of CheXpert labels in 1% train data setting for finetuning of uncased models. Aug-Model correspond to the models pre-trained in this thesis with sentence level data augmentation strategies. Bold and underlined indicates the best result overall.

Model	F1 Macro	F1 Weighted	ROC-AUC Macro	ROC-AUC Weighted
BERT	0.22	0.33	0.67	0.7
SciBERT	0.31	0.49	**0.68**	**0.75**
BioBERT	0.13	0.21	0.58	0.64
Bio-ClinicalBERT	0.1	0.1	0.57	0.58
Aug-Model(Drop)	0.31	**0.53**	0.66	0.72
Aug-Model(Shuffle)	**0.34**	0.52	0.67	**0.75**
Aug-Model(Swap)	**0.34**	**0.53**	**0.68**	**0.75**
Aug-Model(1%-kNeighbors)	0.31	0.52	0.67	**0.75**

Table A.3.: Test set results multi label classification of CheXpert labels in 10% train data setting for finetuning of cased models. Aug-Model correspond to the models pre-trained in this thesis with sentence level data augmentation strategies. Bold and underlined indicates the best result overall.

Model	F1 Macro	F1 Weighted	ROC-AUC Macro	ROC-AUC Weighted
BERT	0.1	0.1	0.51	0.52
SciBERT	0.1	0.18	0.57	0.58
BioBERT	0.1	0.14	0.53	0.54
Bio-ClinicalBERT	0.1	0.1	0.53	0.53
Aug-Model(Drop)	0.14	0.24	0.52	0.53
Aug-Model(Shuffle)	0.1	0.15	**0.58**	**0.59**
Aug-Model(Swap)	**0.15**	0.23	0.56	0.58
Aug-Model(1%-kNeighbors)	0.14	**0.25**	0.56	0.57

Table A.4.: Test set results multi label classification of CheXpert labels in 1% train data setting for finetuning of cased models. Aug-Model correspond to the models pre-trained in this thesis with sentence level data augmentation strategies. Bold and underlined indicates the best result overall.

Model	Accuracy
BERT	0.89
SciBERT	**0.94**
PubmedBERT	0.93
CXR-BERT(General)	0.85
CXR-BERT(Specialized)	0.81
Aug-Model(Drop)	0.93
Aug-Model(Shuffle)	**0.94**
Aug-Model(Swap)	0.93
Aug-Model(5%-kNeighbors)	0.92
Aug-Model(1%-kNeighbors)	0.93

Table A.5.: Accuracy results from named entity recognition of RadGraph dataset for uncased models. Aug-Model models correspond to the pre-trained language models in this thesis. Bold and underlined indicates the best result overall.

Model	Accuracy
BERT	0.89
SciBERT	**0.93**
BioBERT	0.9
Bio-ClinicalBERT	0.89
Aug-Model(Drop)	0.92
Aug-Model(Shuffle)	0.91
Aug-Model(Swap)	0.91
Aug-Model(1%-kNeighbors)	0.92

Table A.6.: Accuracy results from named entity recognition of RadGraph dataset for cased models. Aug-Model models correspond to the pre-trained language models in this thesis. Bold and underlined indicates the best result overall.

For the evaluation of NER task described in Section: 6.4, precision, recall and F1 scores are calculated for the combinations of the four evaluation schemas and five error types described in A.7 and A.8. Examples for the combinations of the errors and evaluation schemas are presented in official documentation [2].

Error Type	Explanation
Correct(COR)	Predicted entity and annotated entity as label are the same.
Incorrect(INC)	Predicted entity and the annotated entity as label don't match.
Partial(PAR)	The model prediction and the annotated entity as label are partially match
Missing(MIS)	The annotated entity is not captured by the model, hence predicted as "*O*"
Spurius(SPU)	The model predicts an entity which doesn't exist annotated entities as labels

Table A.7.: Explanations of Error Types for Named Entity Recognition (NER) of RadGraph Dataset [2].

Evaluation Schema	Explanation
Strict	Exact boundary surface match for tokens and entity type.
Exact	Exact boundary match over the tokens regardless of the entity type.
Partial	Partial boundary match over the tokens, regardless of the entity type.
Type	Some overlap between the predicted entity and the annotated entity as label is required.

Table A.8.: Explanations of Evaluation Schemas for Named Entity Recognition (NER) of RadGraph Dataset [2].

Bibliography

[1] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott. *Publicly Available Clinical BERT Embeddings*. 2019. DOI: 10.48550/ARXIV.1904.03323.

[2] D. Batista and M. A. Upson. *nervaluate*. Version 0.1.8. Oct. 2020.

[3] I. Beltagy, A. Cohan, and K. Lo. "SciBERT: Pretrained Contextualized Embeddings for Scientific Text". In: *CoRR* abs/1903.10676 (2019). arXiv: 1903.10676.

[4] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay. *Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing*. 2022. DOI: 10.48550/ARXIV.2204.09817.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language Models are Few-Shot Learners". In: *CoRR* abs/2005.14165 (2020). arXiv: 2005.14165.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. "A Simple Framework for Contrastive Learning of Visual Representations". In: *CoRR* abs/2002.05709 (2020). arXiv: 2002.05709.

[7] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady. "Structural Scaffolds for Citation Intent Classification in Scientific Publications". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3586–3596. DOI: 10.18653/v1/N19-1361.

[8] X. Dai and H. Adel. "An Analysis of Simple Data Augmentation for Named Entity Recognition". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3861–3867. DOI: 10.18653/v1/2020.coling-main.343.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: `1810.04805`.

[10] H. El Boukkouri, O. Ferret, T. Lavergne, and P. Zweigenbaum. "Re-train or Train from Scratch? Comparing Pre-training Strategies of BERT in the Medical Domain". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 2626–2633.

[11] T. Gao, X. Yao, and D. Chen. "SimCSE: Simple Contrastive Learning of Sentence Embeddings". In: *CoRR* abs/2104.08821 (2021). arXiv: `2104.08821`.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger. Vol. 27. Curran Associates, Inc., 2014.

[13] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *CoRR* abs/2007.15779 (2020). arXiv: `2007.15779`.

[14] B. Gunel, J. Du, A. Conneau, and V. Stoyanov. "Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning". In: *CoRR* abs/2011.01403 (2020). arXiv: `2011.01403`.

[15] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8342–8360. DOI: `10.18653/v1/2020.acl-main.740`.

[16] J. Hong, T. Kim, H. Lim, and J. Choo. "AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain". In: *CoRR* abs/2110.13434 (2021). arXiv: `2110.13434`.

[17] J. Howard and S. Ruder. "Fine-tuned Language Models for Text Classification". In: *CoRR* abs/1801.06146 (2018). arXiv: `1801.06146`.

[18] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence*

Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'19/IAAI'19/EAAI'19. Honolulu, Hawaii, USA: AAAI Press, 2019. ISBN: 978-1-57735-809-1. DOI: `10.1609/aaai.v33i01.3301590`.

[19] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: *CoRR* abs/1901.07031 (2019). arXiv: `1901.07031`.

[20] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3 (May 2016), p. 160035. ISSN: 2052-4463. DOI: `10.1038/sdata.2016.35`.

[21] J. Johnson, M. Douze, and H. Jégou. "Billion-scale similarity search with GPUs". In: *IEEE Transactions on Big Data* 7.3 (2019), pp. 535–547.

[22] P. Kaushik, A. Gain, A. Kortylewski, and A. L. Yuille. "Understanding Catastrophic Forgetting and Remembering in Continual Learning with Optimal Relevance Mapping". In: *CoRR* abs/2102.11343 (2021). arXiv: `2102.11343`.

[23] D. P. Kingma and M. Welling. *Auto-Encoding Variational Bayes*. 2013. DOI: `10.48550/ARXIV.1312.6114`.

[24] F. Koto, J. H. Lau, and T. Baldwin. "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10660–10668. DOI: `10.18653/v1/2021.emnlp-main.833`.

[25] R. Krishnan, P. Rajpurkar, and E. Topol. "Self-supervised learning in medicine and healthcare". In: *Nature Biomedical Engineering* 6 (Aug. 2022), pp. 1–7. DOI: `10.1038/s41551-022-00914-1`.

[26] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. "BioBERT: a pretrained biomedical language representation model for biomedical text mining". In: *CoRR* abs/1901.08746 (2019). arXiv: `1901.08746`.

[27] B. Li, Y. Hou, and W. Che. "Data augmentation approaches in natural language processing: A survey". In: *AI Open* 3 (2022), pp. 71–90. ISSN: 2666-6510. DOI: `https://doi.org/10.1016/j.aiopen.2022.03.001`.

[28] S. Longpre, Y. Wang, and C. DuBois. "How Effective is Task-Agnostic Data Augmentation for Pretrained Transformers?" In: *CoRR* abs/2010.01764 (2020). arXiv: `2010.01764`.

[29] S. Lupart, B. Favre, V. Nikoulina, and S. Ait-Mokhtar. "Zero-Shot and Few-Shot Classification of Biomedical Articles in Context of the COVID-19 Pandemic". In: *CoRR* abs/2201.03017 (2022). arXiv: `2201.03017`.

[30] *MIMIC-CXR Database v2.0.0*. `https://physionet.org/content/mimic-cxr/2.0.0/`.

[31] *MIMIC-CXR Github Repository*. `https://github.com/MIT-LCP/mimic-cxr`.

[32] *MIMIC-CXR-JPG Database v2.0.0*. `https://physionet.org/content/mimic-cxr-jpg/2.0.0/`.

[33] *MIMIC-IV documentation*. `https://mimic.mit.edu/docs/iv/modules/cxr/`.

[34] A. van den Oord, Y. Li, and O. Vinyals. "Representation Learning with Contrastive Predictive Coding". In: *CoRR* abs/1807.03748 (2018). arXiv: `1807.03748`.

[35] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. "Continual lifelong learning with neural networks: A review". In: *Neural Networks* 113 (2019), pp. 54–71. ISSN: 0893-6080. DOI: `https://doi.org/10.1016/j.neunet.2019.01.012`.

[36] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020.

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". In: *CoRR* abs/2103.00020 (2021). arXiv: `2103.00020`.

[38] A. Radford and K. Narasimhan. "Improving Language Understanding by Generative Pre-Training". In: 2018.

[39] *RadGraph: Extracting Clinical Entities and Relations from Radiology Reports*. `https://physionet.org/content/radgraph/1.0.0/`.

[40] *RadNLI: A natural language inference dataset for the radiology domain*. `https://physionet.org/content/radnli-report-inference/1.0.0/`.

[41] L. A. Ramshaw and M. P. Marcus. "Text Chunking using Transformation-Based Learning". In: *CoRR* cmp-lg/9505040 (1995).

[42] C. Rastogi, N. Mofid, and F. Hsiao. "Can We Achieve More with Less? Exploring Data Augmentation for Toxic Comment Classification". In: *CoRR* abs/2007.00875 (2020). arXiv: `2007.00875`.

[43] A. Romanov and C. Shivade. "Lessons from Natural Language Inference in the Clinical Domain". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1586–1596. DOI: `10.18653/v1/D18-1187`.

[44] I. Segura-Bedmar, P. Martınez, and M. Herrero-Zazo. "SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)". In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, June 2013, pp. 341–350.

[45] C. Shivade. *MedNLI - A Natural Language Inference Dataset For The Clinical Domain (version 1.0.0)*. `https://physionet.org/content/mednli/1.0.0/`.

[46] W. Tai, H. T. Kung, X. Dong, M. Comiter, and C.-F. Kuo. "exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1433–1439. DOI: `10.18653/v1/2020.findings-emnlp.129`.

[47] B. Wang, Q. Xie, J. Pei, P. Tiwari, Z. Li, and J. Fu. "Pre-trained Language Models in Biomedical Domain: A Systematic Survey". In: *CoRR* abs/2110.05006 (2021). arXiv: `2110.05006`.

[48] W. Y. Wang and D. Yang. "That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 2557–2563. DOI: `10.18653/v1/D15-1306`.

[49] Z. Wang, Z. Wu, D. Agarwal, and J. Sun. *MedCLIP: Contrastive Learning from Unpaired Medical Images and Text*. 2022. arXiv: `2210.10163 [cs.CV]`.

[50] J. Wei and K. Zou. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6382–6388. DOI: `10.18653/v1/D19-1670`.

[51] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: `10.18653/v1/2020.emnlp-demos.6`.

[52] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: *CoRR* abs/1609.08144 (2016). arXiv: `1609.08144`.

[53] A. Yakimovich, A. Beaugnon, Y. Huang, and E. Ozkirimli. "Labels in a haystack: Approaches beyond supervised learning in biomedical applications". In: *Patterns (New York, N.Y.)* 2.12 (Dec. 2021), p. 100383. ISSN: 2666-3899. DOI: `10.1016/j.patter.2021.100383`.

[54] G. Yan, Y. Li, S. Zhang, and Z. Chen. "Data Augmentation for Deep Learning of Judgment Documents". In: *Intelligence Science and Big Data Engineering. Big Data and Machine Learning*. Ed. by Z. Cui, J. Pan, S. Zhang, L. Xiao, and J. Yang. Cham: Springer International Publishing, 2019, pp. 232–242. ISBN: 978-3-030-36204-1.

[55] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu. *BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model*. 2022. arXiv: `2204.03905 [cs.CL]`.

[56] D. Zhang, T. Li, H. Zhang, and B. Yin. "On Data Augmentation for Extreme Multi-label Classification". In: *CoRR* abs/2009.10778 (2020). arXiv: `2009.10778`.

[57] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. "Contrastive Learning of Medical Visual Representations from Paired Images and Text". In: *CoRR* abs/2010.00747 (2020). arXiv: `2010.00747`.

[58] Z. Zhao, S. Zhu, and K. Yu. "Data Augmentation with Atomic Templates for Spoken Language Understanding". In: *CoRR* abs/1908.10770 (2019). arXiv: `1908.10770`.