Bioinformatics, 36(21), 2020, 5255–5261
doi: 10.1093/bioinformatics/btaa668
Advance Access Publication Date: 23 July 2020
Original Paper

Data and text mining

Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports

Keno K. Bressem [1,2,*,†], Lisa C. Adams [1,2,†], Robert A. Gaudin [3], Daniel Tröltzsch [3], Bernd Hamm [1], Marcus R. Makowski [4], Chan-Yong Schüle [1], Janis L. Vahldiek [1,†] and Stefan M. Niehues [1,†]

[1]Department of Radiology, Charité, Berlin 12203, Germany, [2]Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin 10117, Germany, [3]Department of Oral- and Maxillofacial Surgery, Charité, Berlin 12203, Germany and [4]Department of Diagnostic and Interventional Radiology, Technical University of Munich, School of Medicine, Munich 81675, Germany

*To whom correspondence should be addressed.

[†]The authors wish it to be known that these authors contributed equally.

Associate Editor: Jonathan Wren

Received on May 7, 2020; revised on June 25, 2020; editorial decision on July 14, 2020; accepted on July 17, 2020

Abstract

Motivation: The development of deep, bidirectional transformers such as Bidirectional Encoder Representations from Transformers (BERT) led to an outperformance of several Natural Language Processing (NLP) benchmarks. Especially in radiology, large amounts of free-text data are generated in daily clinical workflow. These report texts could be of particular use for the generation of labels in machine learning, especially for image classification. However, as report texts are mostly unstructured, advanced NLP methods are needed to enable accurate text classification. While neural networks can be used for this purpose, they must first be trained on large amounts of manually labelled data to achieve good results. In contrast, BERT models can be pre-trained on unlabelled data and then only require fine tuning on a small amount of manually labelled data to achieve even better results.

Results: Using BERT to identify the most important findings in intensive care chest radiograph reports, we achieve areas under the receiver operation characteristics curve of 0.98 for congestion, 0.97 for effusion, 0.97 for consolidation and 0.99 for pneumothorax, surpassing the accuracy of previous approaches with comparatively little annotation effort. Our approach could therefore help to improve information extraction from free-text medical reports.

Availability and implementation

We make the source code for fine-tuning the BERT-models freely available at https://github.com/fast-raidiology/bert-for-radiology.

Contact: keno-kyrill.bressem@charite.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Over the past years, deep learning has fundamentally changed the landscapes of numerous areas in artificial intelligence, including natural language processing (NLP). NLP is dedicated to the building of computational algorithms for automated analysis and representation of the human language. Development of pre-trained language representation models has vastly improved many NLP applications (Devlin *et al.*, 2018; Howard and Ruder, 2018; Peters *et al.*, 2018; Tenney *et al.*, 2019). In particular, the development of Bidirectional Encoder Representations from Transformers (BERT) substantially outperformed several benchmarks in multiple common NLP tasks (Devlin *et al.*, 2018). The superior performance of BERT models is due to several factors: first, BERT models are based on transfer learning, which means that the core model is pre-trained on large amounts [several gigabytes (GB)] of unlabelled texts to gain knowledge of text structure and patterns of language. Since only a limited amount of labelled data is available (up to several thousand labelled texts versus up to several billions of freely available unlabelled texts, e.g. Wikipedia), BERT models enable substantially better understanding of language than other deep learning models in NLP, such as recurrent neural networks (RNN) (Goldberg, 2019; Vaswani *et al.*, 2017). Therefore, fine-tuned BERT models achieve better results than conventional methods, at the same time requiring less labelled data. Second, the model is deeply bidirectional, which means it can recognize the meaning of a word in a sentence based on the words before and after the word. This distinguishes BERT from previous similar models, which remained mostly unidirectional or shallowly bidirectional (Devlin *et al.*, 2018; Peters *et al.*, 2018; Radford *et al.*, 2018). Using the transfer learning approach, BERT models can be pre-trained locally and then made publicly available, so that others may fine-tune them for specific tasks, such as text-based label extraction for image classification, without having to repeat pre-training.

Given the potential of the BERT models, their transfer to medical research appears promising, although the large domain-specific vocabulary would likely pose a challenge. Especially radiology departments generate large quantities of digital text reports. The use of these data for healthcare research would be highly beneficial, as they are annotated by medical experts and could thus be used to perform large-scale research projects with the prospect to improve clinical care. Although text reports are stored for documentation of diagnostic imaging, utilizing their potential requires new forms of automated information extraction as the data exist mainly as narrative free text. So far, mostly conventional NLP models, such as rule-based algorithms, RNN or convolutional neural networks, have been applied to radiology reports (Cai *et al.*, 2016, 2018; LeCun *et al.*, 2015; Pons *et al.*, 2016). A challenge in developing these algorithms is that time-consuming labelling of text reports by human experts is required. Consequently, large amounts of annotated data are usually hard to come by, especially for non-English reports (Hosny *et al.*, 2018; Pinto Dos Santos *et al.*, 2019).

To overcome these challenges, we propose four different BERT models, of which two were pre-trained on 3.8 million radiology free-text reports from our institution, and two were only based on open-source models. We first fine-tuned the models, so that they would automatically recognize and sort out texts without sufficient information for the extraction of labels and then fine-tuned for label extraction from radiology reports of chest radiographs as well as chest CT scans from the intensive care unit. Subsequently, we compared the performance of the four proposed models.

2 Materials and methods

In this study, the performance of four models in classifying free-text chest radiograph reports performed in the intensive care unit was compared against each other. For all four models, the BERT-base architecture was used, which consisted of 12 layers, 12 heads and 110 million parameters (Devlin *et al.*, 2018). The four models included a German BERT model (GER-BERT) published by deepset (deepset GmbH, Berlin, Germany) and the multilingual cased BERT model (MULTI-BERT) published by Devlin *et al.*, and both models were only fine-tuned on the data for the classification task (Devlin *et al.*, 2018). Furthermore, we trained a BERT model from scratch (FS-BERT), solely based on our text corpus, which consisted of 3.8 million unstructured radiology reports. The fourth model was a BERT model based on GER-BERT, which underwent additional pre-training steps on our domain-specific text corpus of radiology text reports (RAD-BERT).

The code and a detailed documentation of all steps described in this article can be accessed on the accompanying GitHub repository (github.com/fast-raidiology/bert-for-radiology). Modelling was mainly performed using Python (version 3.7) (Van Rossum and Drake, 2009. This study was approved by the institutional review board (EA4/042/20) in advance.

2.1 Text extraction

All free-text reports generated between January 1, 2009 and June 1, 2019 were extracted ($n = 4\ 790\ 000$) from our institution's radiological information system (RIS), which contained reports from radiographs, computed tomography, magnetic resonance imaging, ultrasound, nuclear medicine, radiation therapy and interventional radiology. Out of these, $n = 948\ 457$ were removed, as they did not contain chest radiograph reports (i.e. constancy testing). To identify these reports, we sorted all diagnostic texts according to how frequently their exact character combination occurred. Since the exact wording of free texts is highly unlikely to occur multiple times, texts which are automatically filled in, e.g. about the non-appearance of patients or image imports could be reliably identified and then excluded. Also, texts with a character count of less than 100 were excluded, leaving $n = 3\ 841\ 543$ reports for pre-training (415 702 033 words, 3.36 GB). After white-space stripping, the reports were divided into their individual sentences using the 'spaCy' library for Python and stored as a single raw text file (Honnibal and Montani, 2017).

2.2 Data preparation

As radiologic texts contain numerous domain-specific words, which were not present in the vocabulary files of the GER-BERT or MULTI-BERT models, we created a custom WordPiece vocabulary for pre-training our own models (RAD-BERT, FS-BERT). As the official BERT repository does not contain code to generate new vocabulary, we used the open source 'bert-vocab-builder' by Kwon (Kwon, 2019). Words appearing more than 1000 times within our corpus were included, resulting in a vocabulary size of 29 502 (sub)-words, which was then padded to a size of 30 000 to match the size defined in the configuration file of GER-BERT.

Based on the custom WordPiece vocabulary, raw texts could be transformed into a format suitable for training BERT models using the open-source scripts of the Google AI Research team from the official BERT GitHub repository (github.com/google-research/bert). This data generation was performed twice with the maximum sequence length set to 128 and 512 tokens and a masked words percentage of 15% (whole word masking). Due to memory issues, the raw texts had to be split into smaller files containing one million lines of text each, resulting in 55 training files for each of the two sequence lengths. Generation of training data as well as the next steps (pre-training and fine-tuning) were carried out using the TensorFlow library (Abadi *et al.*, 2016).

2.3 Pre-training

Pre-training of BERT models includes two tasks: first, prediction of the words that were previously masked in the text corpus [masked language models (LM) accuracy], enabling the model to understand contextual word embeddings. Second, next sentence prediction, a task in which the model must predict whether the second sentence is likely to follow the first, through which the model can learn relationships between sentences (Devlin *et al.*, 2018). We pre-trained two different BERT models. The first model was trained from scratch using only our text data (FS-BERT). For the second model, we used the open-source German BERT model as initial checkpoint (RAD-BERT). For both models, 90 000 training steps including 9000 warmup steps were executed for a maximum sequence length of 128 tokens and an additional 10 000 training steps including 1000 warmup steps for a maximum sequence length of 512 tokens. Our corpus contained 415 702 033 words, with the number of training steps corresponding to approximately 40 epochs. The batch size was 32 for 128 tokens and 6 for 512 tokens. The learning rate was kept constant at $2e^{-5}$. The models were then converted to a format useable by the PyTorch library, utilizing the Hugging Face's Transformers library (Paszke *et al.*, 2017; Wolf *et al.*, 2019).

2.4 Annotation

About 7200 text reports of bedside chest radiographs were randomly extracted and manually annotated by four experienced radiologists (KKB, LCA, SMN and JLV) and an oral and maxillofacial surgeon (RAG). Annotations were performed with an in-house developed tool (annotator), which was specifically programmed for this purpose. This tool can also be accessed on the GitHub repository accompanying this article (github.com/fast-raidiology/bert-for-radiology). The texts were independently annotated twice and, if annotations contradicted each other, a third annotation was performed by a third observer, who was blinded to the previous annotations. Inclusion and exclusion criteria for the annotations are given in the Appendix. The final annotation was the one made by the majority of evaluating radiologists. Nine findings were annotated: congestion, opacity (e.g. pneumonia, dystelectasis), effusion, pneumothorax, central venous catheters, gastric tube, thoracic drains, tracheal tube (or cannula) and misplaced medical device. From a clinical point of view, we considered the combination of pneumonia, dystelectasis and other airway consolidations into one label to be justifiable since exact differentiation of pneumonia and dystelectasis is not always possible based on a chest radiograph alone or the written report, and clinical information need to be considered as well. If the text did not contain enough information, for example only reported absence of pneumothorax after insertion of a venous catheter without any further information, it was marked as incomplete ($n = 1997$), allowing subsequent training of a binary classifier for automated pre-selection of reports.

The annotated dataset was then randomly split into a test (500 reports) and a training (4703 reports) dataset for the subsequent fine-tuning tasks. Cohen's kappa (κ) was used to assess inter-rater agreement between radiologists. Example texts and annotations can be found in Supplementary Table S6.

2.5 Pre-selection of text reports

Free texts without sufficient information for classification with our pre-defined labels were marked as incomplete (also refer to the previous paragraph). Since it is not possible to extrapolate the presence or absence of any undisclosed findings, we decided to exclude these reports, because otherwise the labels generated from these texts would be inaccurate and might bias downstream tasks depending on those labels. For this task, all four models were fine-tuned as binary classifiers on the annotated data with tuning parameters held constant for all models. We used a maximum sequence length of 512 tokens, a training batch size of 8 (6 for the multi-lingual model due to out-of-memory errors) and a learning rate of $4e^{-5}$. The training of one binary classifier took approximately 20 min on two GeForce RTX 2080 Ti graphic cards (Nvidia Corporation, Santa Clara, California, USA).

2.6 Fine-tuning for multi-label classification

Fine-tuning was done using the Simple Transformers library, which is built on top of the Hugging Face Transformers library (Rajapakse, 2019; Wolf *et al.*, 2019). The four models were refined for multi-label classification with tuning parameters held constant for all models. Hyperparameters were the same as for the pre-selection task. Fine-tuning was executed with two GeForce RTX 2080 Ti graphic cards and took approximately 90 min. We trained for 32 epochs, and the loss and label ranking average precision scores (LRAP) on the evaluation dataset were plotted against the numbers of epochs (see Supplementary Fig. S1). From this, the optimal number of training epochs was determined.

To determine the gain in accuracy achieved by increasing the amount of training data, we performed repeated fine-tuning using the pre-defined optimal number of training epochs as defined before and increasing amounts of training data ranging from a size of 100 to 4500 texts, randomly drawn from the total training data. The final models were then fine-tuned on to the entire train dataset.

2.7 Classification of CT reports

Since the report texts for chest radiographs are short and we assumed that the benefit of pre-training and the custom WordPiece vocabulary might only become apparent with longer texts, we additionally evaluated all trained models on CT reports. For this, we identified CT reports on examinations that were performed within 24 h before or after a chest radiograph. Then 100 reports were randomly selected and triple-annotated. Assuming that wordings concerning the presence or absence of the above-specified findings would not differ substantially between CT and chest radiograph text reports, we did not perform additional training on CT reports. As the maximum processing capacity of our models was fixed at 512 sub-word tokens and consequently texts exceeding this threshold would have been lost, we used a sliding window approach with 20% overlap, where each 512-token paragraph was evaluated independently with an averaging of predictions. In case of a tie, the models assessed the finding as positive.

2.8 Evaluation of model performance

The raw predictions of the model were exported as comma-separated values and further processed using the statistical language 'R' (version 3.6) and the 'tidyverse' library. The raw predictions could be values between -infinity and infinity, which were scaled to values between 0 and 1 using a softmax function. To transform these predictions into either 1 (finding predicted as present) or 0 (finding predicted as absent) we used a threshold of 0.5, with values smaller than 0.5 indicating that the finding was not present. This threshold was always kept constant, and no attempt was made to achieve an increase in accuracy on the test data by changing the threshold.

Different scores were used to evaluate and compare the four models. To plot the increase in accuracy with increasing training size, we chose the pooled F1 score and accuracy score. For a more detailed overview of model performance on selected sizes of the train dataset, we used radar plots of the F1 score, the Youden index (sensitivity + specificity - 1) and Matthews correlation coefficient (MCC). The MCC was used as it considers all values derived from a confusion matrix and remains reliable even with strong class imbalances. The final models were then evaluated using the above-mentioned scores as well as the area under the receiver operating characteristic curve (AUC) and the area under the precision–recall curve (AUPRC).

2.9 Code availability

All code as well as the pre-trained TensorFlow model and PyTorch model of RAD-BERT can be found at github.com/fast-raidiology/bert-for-radiology.

3 Results

After removing incomplete reports, annotations for 5783 text reports were made by four radiologists. Among those texts, 6% ($n = 303$) were annotated as normal with no pathologic findings and no medical devices inserted, and 19% ($n = 992$) reported no pathologic findings, but at least one inserted medical device. Opacity was the most common pathologic finding with a prevalence of 60% ($n = 3101$), followed by effusion with a prevalence of 47% ($n = 2461$) and congestion with a prevalence of 28% ($n = 1446$). Pneumothorax was the least frequent finding and was reported in only 8% ($n = 429$) of annotated report texts, leading to a strong class imbalance for this finding.

The pooled percentage agreement between the radiologists was 93.7%. Accuracy by finding was 91.8% ($n = 5309$) for congestion, 83.2% ($n = 4813$) for opacity, 85.8% ($n = 4962$) for effusion and 97.2% ($n = 5619$) for pneumothorax. κ, as a measurement of inter-rater agreement, was 0.77 for congestion, 0.67 for opacity, 0.71 for effusion and 0.83 for pneumothorax, which can be interpreted as moderate inter-rater agreement (McHugh, 2012).

The most common medical devices were venous catheters with a prevalence of 58% ($n = 3020$), followed by tracheal tubes with a prevalence of 41% ($n = 2110$) and gastric tubes with a prevalence of 25% ($n = 1299$). Thoracic drain was the least common medical device with a prevalence of 21% ($n = 1077$). Inter-rater agreement was excellent for all medical devices with an agreement of 96.6% ($\kappa = 0.93$), 97.3% ($\kappa = 0.94$), 97.8% ($\kappa = 0.94$) and 96.2% ($\kappa = 0.9$). Malposition of a medical device was reported in 4% ($n = 188$) of all reports, resulting in a class imbalance for this finding, which is also reflected in the overall poor accuracy of the models in detecting reported malposition.

3.1 Co-occurrence of findings

The co-occurrences of the different findings were similar for the train/valid and the test dataset. Venous catheters co-occurred the most frequently with other medical devices as well as pathophysiological findings. Furthermore, 64% (train/valid) to 67% (test) reports of radiographs mentioning pneumothorax also mentioned a thoracic drain. Results on co-occurrences are visualized in Supplementary Figure S2.

3.2 Pre-training accuracy

Pre-training was carried out on a single GeForce RTX 2080 Ti graphics card with 11 GB or RAM and took approximately 2 days. After pre-training, the model trained with the German BERT-base as initial checkpoint showed a masked LM accuracy of 69.9% and a next sentence accuracy of 94.8%. The BERT model pre-trained only on the domain-specific corpus showed a masked LM accuracy of 65.1% and a next sentence accuracy of 89.1%.

3.3 Pre-selection of report texts

Fine-tuning a BERT model as binary classifier took approximately 15 min for each model. All BERT models showed a very high accuracy for detecting unevaluable reports. Both RAD-BERT and GER-BERT achieved F1 scores of 0.98 followed by MULTI-BERT and FS-BERT, which both achieved a F1 score of 0.97. So, 1.6–2.3% of all texts were wrongly labelled as unevaluable by RAD-BERT and GER-BERT.

3.4 Classification accuracy for chest radiograph reports

The best performance of the models was achieved after four epochs of fine-tuning, as can be seen in Supplementary Figure S1. We therefore used these checkpoints to evaluate performance in classification. We performed fine-tuning on different sizes of training data to determine the gain in performance through the use of increasingly larger datasets. The results show that no further significant improvements were achieved when using a training dataset of 2000 texts or above (see Supplementary Figs S2 and S3). With a smaller training dataset (200–1000 texts), our pre-trained model, based on the German BERT model, showed a better performance than the German BERT model, the multi-lingual model or the BERT model pre-trained from scratch only on the domain-specific corpus. Especially for the detection of pneumothorax, which showed a strong class imbalance, this model performed better. Using a train dataset of 1000 annotated texts for fine-tuning, we achieved the highest accuracy with our pre-trained model with a pooled AUC of 0.98 and an AUPRC of 0.89, followed by the MULTI-BERT model with a pooled AUC and AUPRC of 0.97 and 0.86, respectively, and the GER-BERT model (AUC 0.96, AUPRC 0.83).

The poorest performance was achieved by the BERT model trained from scratch only on the domain-specific corpus with a pooled AUC/AUPRC of 0.95/0.75. As the size of the dataset increased (≥2000 texts), the differences in performance diminished (see Supplementary Figs S3 and S4). The best overall accuracy was achieved with a training dataset size of 4000 texts. The pooled AUC/AUPRC of the models fine-tuned on 4000 text reports was 0.98/0.93 for our RAD-BERT, GER-BERT as well as MULTI-BERT. FS-BERT showed a poorer performance with an AUC of 0.98 and an AUPRC of 0.91. The F1 score, Youden's index (sensitivity + specificity – 1) and MCC for each model (using a standard threshold of 0.5 to distinguish between a positive and negative finding), fine-tuned on different-sized training data, are provided in Supplementary Tables S1 and S2. AUC und AUPRC values for different-sized datasets are given in Supplementary Table S3 and are displayed in Figure 1. Please refer to Supplementary Table S5 for a comparison with previously published algorithms. The results are also presented in more detail in the accompanying GitHub repository.

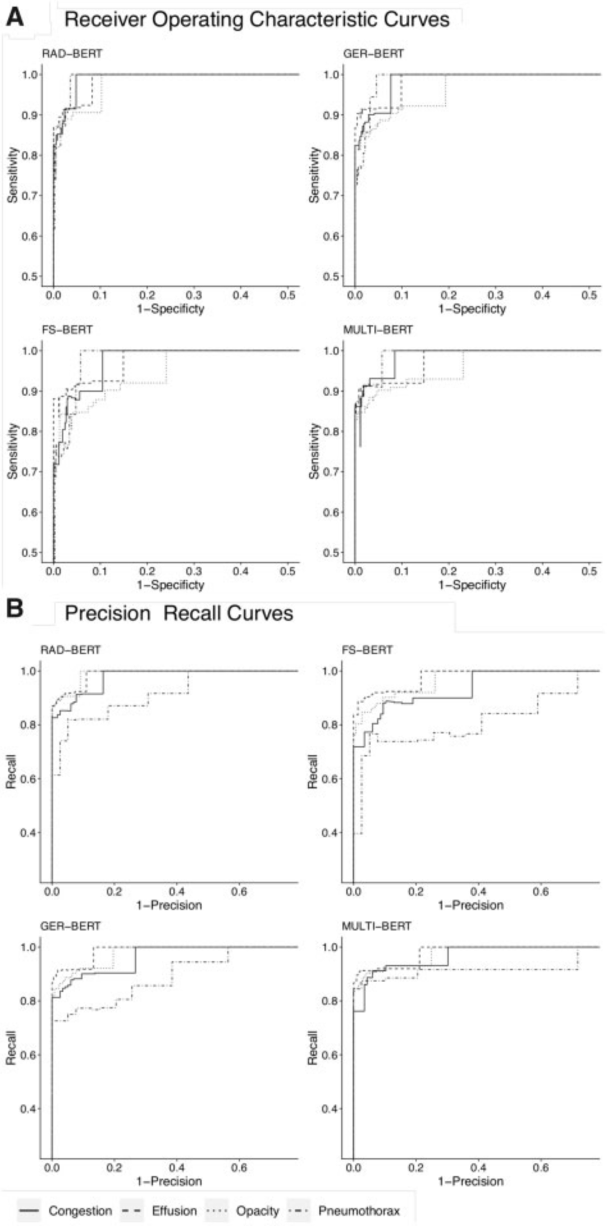

Fig. 1. This shows the receiver operating characteristic (ROC) curves (**A**) and precision–recall (PR) curves (**B**) of the different models for congestion, effusion, opacity and pneumothorax. As the curves are close to each other, the axis of the plot has been zoomed to 0.5–1 (y) and 0–0.5 (x) for A and to 0.25–1 (y) and 0–0.75 (x) for B. While the ROC curves and PR curves for GER-BERT, MULTI-BERT and RAD-BERT are similar, FS-BERT shows a slightly poorer performance in the PR curves, especially for pneumothorax. Possible reasons include that FS-BERT was only pre-trained on the radiological corpus and no additional texts and that there was a class imbalance for pneumothorax

3.5 Performance on CT reports

Since diagnostic reports on chest radiographs are relatively short, we also evaluated the models on CT text reports. We selected 100 CT examinations performed within 24 h before or after a chest radiograph which either reported only on the chest ($n = 37$) or also contained texts on the head ($n = 43$), neck ($n = 17$) or abdomen ($n = 62$). With an average word count of 240 (SD 110, min 55, max 592), CT reports were substantially longer than the chest radiograph reports with a mean word count of 66 (SD 20, min 22, max 155). Without additional training, we used the four BERT models for text classification. This showed a superiority of our RAD-BERT model over the other BERT models with a pooled AUC/AUPRC of 0.88/

0.80, followed by the MULTI-BERT model (pooled AUC/AUPRC = 0.87/0.75) and the GER-BERT model (pooled AUC/AUPRC = 0.85/0.74). RO-BERT showed the poorest performance with a pooled AUC/AUPRC of 0.78/0.62. Further details of AUC and AUPRC results are compiled in Supplementary Table S4.

4 Discussion

BERT models, such as our in-house developed RAD-BERT, enable a fast and accurate classification of large amounts of radiology reports, such as chest radiograph or CT reports. At the same time, they surpass previously published algorithms, while only requiring little annotation effort. Consequently, the implementation of our RAD-BERT model for other radiology-related NLP tasks would be very time-efficient, as the annotation of 1000 texts, which can be accomplished within 10 h, was sufficient to achieve state-of-the-art results. Our results also show that additional pre-training on a domain-specific word corpus can speed up the fine-tuning task by reducing the overall amount of labelled data needed.

We believe that the use of BERT models can help to improve and accelerate further tasks, especially in machine learning, by providing highly accurate text-based labels for image classification. Moreover, the models might also be used for structured reporting or in a system to immediately notify the treating physician whenever an important finding is mentioned in a radiologist's report.

All our code and pre-trained models are provided on our GitHub repository to enable other researchers to implement these models for their domain-specific tasks (github.com/fast-raidiology/bert-for-radiology).

There have been previous studies, applying different NLP tools to radiology reports, with the majority focusing on English text reports. Even though a large part of the world speaks English, in most countries, free-text reports are written in the native language, precluding the use of already published algorithms or of English-language libraries containing domain-specific words such as RadLex (Langlotz, 2006). By contrast, the approach presented in this study enables the use of a BERT model pre-trained in the respective language to learn the domain-specific words or of a multilingual BERT model as groundwork, which may then be fine-tuned to the domain-specific text corpus based on transfer learning.

The strength of transfer learning is that only the core BERT model needs to be adapted for implementation in another language, while the rest of the code remains unchanged. Conversely, rule-based algorithms have to be fundamentally changed to adapt them to the specific characteristics of the respective language. Previous studies also dealt with the classification of free text reports, especially reports on chest radiographs, and were mostly outperformed by our model (summarized in Supplementary Table S5). For example, Elkin *et al.* implemented a rule-based algorithm to extract pneumonia cases from text reports, yet they did not extract further findings. The disadvantage of their algorithm is that it will only work on the given task, while every new application, e.g. extraction of pneumothorax findings, will require the development and evaluation of a new algorithm. In contrast, when using BERT, only the fine-tuning has to be adapted for a novel task.

Table 1. Inter-rater agreement

Finding	Kappa value	Agreement
Congestion	0.77*	91.8%
Opacity	0.67*	83.2%
Effusion	0.71*	85.8%
Pneumothorax	0.83*	97.2%
Thoracic drain	0.90*	96.2%
Venous catheter	0.93*	96.6%
Gastric tube	0.94*	97.8%
Tracheal tube	0.94*	97.3%
Misplaced medical device	0.71*	97.5%
Pooled	0.82*	93.7%

Note: Table 1 gives an overview of the inter-rater agreement between the human annotators for different findings. It can be seen that even in the classification of report texts, there can be disagreements of up to 17%, as for example with the annotation of opacity, which results from the partly vague language expressions used by radiologists.

*$P < 0.001$.

Chen *et al.* used a deep learning convolution neural network (CNN) to extract pulmonary embolism findings from radiology reports and surpassed previous tools such as PeFinder. Although they also utilized a shallow unsupervised algorithm to generate word embeddings, called GloVe (Global Vectors for Word Representation), this is different from contextual embeddings as used in the deep learning approach by BERT. The difference is that GloVe creates a numeric representation for each word, while the embeddings derived from BERT differ depending on the context. Consequently, BERT will use different numeric representations for homonyms, while GloVe only applies one representation.

Other researchers applied rule-based algorithms to extract labels for radiological images from texts, which were then used to develop image classification models (Bustos *et al.*, 2019; Irvin *et al.*, 2019). In this context, incomplete or very short texts represent an important challenge. For example, in some reports, only a specific pathology was evaluated (e.g. 'No pneumothorax. Otherwise no changes.'), while other reports only mentioned medical devices as unchanged without describing which specific devices were present. If such reports are used to extract labels, e.g. for images, there is a risk of bias. As a high quality of data is essential for the development of reliable machine learning algorithms, biased labels might lead to a poorer performance of the derived models (Gianfrancesco *et al.*,

Table 2. Distribution of findings

Finding	Prevalence
All reports	
Opacity	60% ($n = 3101$)
Effusion	47% ($n = 2461$)
Congestion	28% ($n = 1446$)
Pneumothorax	8% ($n = 429$)
Venous catheter	58% ($n = 3020$)
Tracheal tube/cannula	41% ($n = 2110$)
Gastric tube	25% ($n = 1299$)
Thoracic drain	21% ($n = 1077$)
Misplaced medical device	4% ($n = 188$)
Train dataset	
Opacity	60% ($n = 2836$)
Effusion	48% ($n = 2243$)
Congestion	28% ($n = 1330$)
Pneumothorax	8% ($n = 390$)
Venous catheter	58% ($n = 2732$)
Tracheal tube/cannula	40% ($n = 1902$)
Gastric tube	25% ($n = 1169$)
Thoracic drain	21% ($n = 969$)
Misplaced medical device	3% ($n = 164$)
Test dataset	
Opacity	53% ($n = 265$)
Effusion	44% ($n = 218$)
Congestion	23% ($n = 116$)
Pneumothorax	8% ($n = 39$)
Venous catheter	58% ($n = 288$)
Tracheal tube/cannula	42% ($n = 208$)
Gastric tube	26% ($n = 130$)
Thoracic drain	22% ($n = 108$)
Misplaced medical device	5% ($n = 24$)

Note: Table 2 gives an overview of the distribution of findings across the overall dataset, the train data and the test data. Specifically, pneumothorax and misplaced medical device showed a strong class imbalance.

2018). To avoid such bias in labels obtained from our BERT models, we first trained the models to identify critical texts based on our annotations, so that they could be excluded if the models were used to classify large amounts of unlabelled data. The models were then fine-tuned for the classification of the chest radiograph reports.

In radiology, many domain-specific words are used, which are unknown to BERT models trained on openly available texts. As a result, we pre-trained two BERT models on a radiology-specific word corpus and created a domain-specific WordPiece Vocabulary, which allowed us to improve model performance with less training data or longer texts. Similar to other domain-specific BERT implementations, the improved performance might be due to the domain-specific WordPiece vocabulary, enabling more efficient tokenization (Beltagy *et al.*, 2019; Lee *et al.*, 2019). As BERT can only process a token length of up to 512 (a token represents a word or sub-word), information may be lost in longer texts. We introduced a custom WordPiece vocabulary, containing domain-specific words such as ‘pneumothorax’, which therefore do not need to be split into sub-word tokens by our models (RAD-BERT, FS-BERT). The GER-BERT model splits it into six tokens (‘*P*’, ‘##ne’, ‘##um’, ‘##ot’, ‘##hor’, ‘##ax’; with ‘##’ indicating that this token is a sub-word token). This reduces the number of words the model can process at one time. Although one can use a sliding window method for longer texts, models with inefficient vocabulary need to evaluate more windows, which carry the risk of information loss if the results are averaged in the end. However, to implement a BERT model for relatively short text reports, increasing the amount of data for fine-tuning seemed to be more important than the custom WordPiece vocabulary or additional pre-training, as the GER-BERT or MULTI-BERT yielded the same accuracy as RAD-BERT. The overall poorer performance of the FS-BERT model could have resulted from the fact that the domain-specific corpus alone might be insufficient to learn proper word embeddings.

Our study has some limitations. Due to memory limitations of the hardware accelerator we used, only the base model could be trained. Devlin *et al.* proposed a larger BERT model, which requires dedicated hardware in the form of TPU, e.g. available via Google-cloud. However, cloud export of sensitive patient data, such as radiology reports, is often not in accordance with local data protection guidelines (Devlin *et al.*, 2018). While a larger BERT model might have yielded a better performance than our base model, it could, in contrast, also have been at risk to remain undertrained on our text corpus.

Future studies should investigate an approach that allows the integration of texts, in which only one finding is reported. In the present analysis, we excluded texts if they only reported one finding and referred to the others as unchanged. Through matching those reports with previous reports one could copy those labels. However, as we anonymized our texts, retrospective matching was not possible.

5 Conclusion

In this study, we successfully used highly accurate and robust BERT models for the classification of radiology text reports, surpassing the accuracy of previous approaches with comparatively little annotation effort. This approach could also easily be transferred to other medical disciplines, empowering physicians and researchers to analyse and classify large amounts of domain-specific texts without the need to develop novel algorithms for each new application and to then use them for further downstream tasks.

Funding

L.C.A. is grateful for her participation in the BIH Charité – Junior Clinician and Clinician Scientist Program, and K.K.B. is grateful for his participation in the BIH Charité Digital Clinician Scientist Program, both funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health. M.R.M. is grateful for support from the Deutsche Forschungsgemeinschaft (DFG, SFB 1340/1 2018, 5943/31/41/91).

Conflicts of Interest

B.H. has received research grants for the Department of Radiology, Charité-Universitätsmedizin Berlin, from the following companies and institutions: (1) Abbott, (2) Actelion Pharmaceuticals, (3) Bayer Schering Pharma, (4) Bayer Vital, (5) BRACCO Group, (6) Bristol-Myers Squibb, (7) Charité Research Organisation GmbH, (8) Deutsche Krebshilfe, (9) Dt. Stiftung für Herzforschung, (10) Essex Pharma, (11) EU Programmes, (12) FibrexMedical Inc., (13) Focused Ultrasound Surgery Foundation, (14) Fraunhofer Gesellschaft, (15) Guerbet, (16) INC Research, (17) lnSightec Ud, (18) IPSEN Pharma, (19) Kendlel MorphoSys AG, (20) Lilly GmbH, (21) Lundbeck GmbH, (22) MeVis Medical Solutions AG, (23) Nexus Oncology, (24) Novartis, (25) Parexel Clinical Research Organisation Service, (26) Perceptive, (27) Pfizer GmbH, (28) Philipps, (29) Sanofis-Aventis S.A., (30) Siemens, (31) Spectranetics GmbH, (32) Terumo Medical Corporation, (33) TNS Healthcare GMbH, (34) Toshiba, (35) UCB Pharma, (36) Wyeth Pharma, (37) Zukunftsfond Berlin (TSB), (38) Amgen, (39) AO Foundation, (40) BARD, (41) BBraun, (42) Boehring Ingelheimer, (43) Brainsgate, (44) PPD (Clinical Research Organisation), (45) CELLACT Pharma, (46) Celgene, (47) CeloNova Bio-Sciences, (48) Covance, (49) DC Deviees, Ine. USA, (50) Ganymed, (51) Gilead Sciences, (52) GlaxoSmithKline, (53) ICON (Clinical Research Organisation), (54) Jansen, (55) LUX Bioseienees, (56) MedPass, (57) Merek, (58) Mologen, (59) Nuvisan, (60) Pluristem, (61) Quintiles, (62) Roehe, (63) SehumaeherGmbH (Sponsoring eines Workshops), (64) Seattle Geneties, (65) Symphogen, (66) TauRx Therapeuties Ud, (67) Accovion, (68) AIO: Arbeitsgemeinschaft Internistische Onkologie, (69) ASR Advanced sleep research, (70) Astellas, (71) Theradex, (72) Galena Biopharma, (73) Chiltern, (74) PRAint, (75) lnspiremd, (76) Medronic, (77) Respicardia, (78) Silena Therapeutics, (79) Spectrum Pharmaceuticals, (80) St Jude, (81) TEVA, (82) Theorem, (83) Abbvie, (84) Aesculap, (85) Biotronik, (86) Inventivhealth, (87) ISATherapeutics, (88) LYSARC, (89) MSD, (90) Novocure, (91) Ockham Oncology, (92) Premier-Research, (93) Psi-cro, (94) Tetec-ag, (95) Winicker-Norimed, (96) Achaogen Inc., (97) ADIR, (98) AstraZenaca AB, (99) Demira Inc., (100) Euroscreen S.A., (101) Galmed Research and Development Ltd, (102) GETNE, (103) Guidant Europe NV, (104) Holaira Inc., (105) Immunomedics Inc., (106) Innate Pharma, (107) Isis Pharmaceuticals Inc., (108) Kantar Health GmbH, (109) MedImmune Inc., (110) Medpace Germany GmbH (CRO), (111) Merrimack Pharmaceuticals Inc., (112) Millenium Pharmaceuticals Inc., (113) Orion Corporation Orion Pharma, (114) Pharmacyclics Inc., (115) PIQUR Therapeutics Ltd, (116) Pulmonx International Sárl, (117) Servier (CRO), (118) SGS Life Science Services (CRO) and (119) Treshold Pharmaceuticals Inc. The funding had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The remaining authors declare that they have no conflicts of interest and did not receive any funds. There are no patents, products in development, or marketed products to declare.

References

Abadi, M. et al. (2016) Tensorflow: a system for large-scale machine learning. In, 12thUSENIX Symposium on Operating Systems Design and Implementation (USENIX association, Savannah, GA, USA). pp. 265–283.

Beltagy,I. *et al.* (2019) SciBERT: A pretrained language model for scientific text. 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)*. pp. 3606–3611.

Bustos,A. *et al.* (2019) Padchest: a large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441*.

Cai,T. *et al.* (2016) Natural language processing technologies in radiology research and clinical applications. *RadioGraphics*, **36**, 176–191.

Chen,M.C. *et al.* (2018) Deep learning to classify radiology free-text reports. *Radiology*, **286**, 845–852.

Devlin,J. *et al.* (2018) Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Honnibal,M. and Montani,I. (2017) spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. URL: https://spacy.io.

Gianfrancesco,M.A. *et al.* (2018) Potential biases in machine learning algorithms using electronic health record data. *JAMA Int. Med.*, **178**, 1544–1547.

Goldberg,Y. (2019) Assessing BERT’s syntactic abilities. *arXiv preprint arXiv: 1901.05287*.

Hosny,A. *et al.* (2018) Artificial intelligence in radiology. *Nat. Rev. Cancer*, **18**, 500–510.

Howard,J. and Ruder,S. (2018) Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Irvin,J. *et al.* (2019) Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*.

Kwon,M.H. (2019) *Bert-Vocab-Builder*. GitHub.

Langlotz,C.P. (2006) *RadLex: A New Method for Indexing Online Educational Materials*. Radiological Society of North America.

LeCun,Y. *et al.* (2015) Deep learning. *Nature*, **521**, 436–444.

Lee,J. *et al.* (2019) Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv: 1901.08746*.

McHugh,M.L. (2012) Interrater reliability: the kappa statistic. *Biochem. Med. (Zagreb)*, **22**, 276–282.

Paszke,A. *et al.* (2019) Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems (NeurIPS, Vancouver, Canada). pp. 8026-8037.

Peters,M.E. *et al.* (2018) Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pinto Dos Santos,D. *et al.* (2019) Structured report data can be used to develop deep learning algorithms: a proof of concept in ankle radiographs. *Insights Imaging*, **10**, 93–93.

Pons,E. *et al.* (2016) Natural language processing in radiology: a systematic review. *Radiology*, **279**, 329–343.

Van Rossum,G. and Drake,F.L. (2009) Python 3 Reference Manual, Scotts Valley, CA: CreateSpace

Radford,A. *et al.* (2018) Improving language understanding with unsupervised learning. In.: *Technical Report, OpenAI*.

Rajapakse,T. (2019) Simple Transformers. GitHub repository, https://github.com/ThilinaRajapakse/simpletransformers.

Tenney,I. *et al.* (2019) Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Vaswani,A. *et al.* (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems, NeurIPS*, Long Beach, CA, USA. pp. 5998–6008.

Wolf,T. *et al.* (2019) Transformers: state-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.