A Multi-Task Transformer Model for Fine-grained Labelling of Chest X-Ray Reports

Yuanyi Zhu, Maria Liakata, Giovanni Montana
University of Warwick, Queen Mary University of London
yuanyi.zhu@warwick.ac.uk, m.liakata@qmul.ac.uk, g.montana@warwick.ac.uk

Abstract

Precise understanding of free-text radiology reports through localised extraction of clinical findings can enhance medical imaging applications like computer-aided diagnosis. We present a new task, that of segmenting radiology reports into topically meaningful passages (segments) and a transformer-based model that both segments reports into semantically coherent segments and classifies each segment using a set of 37 radiological abnormalities, thus enabling fine-grained analysis. This contrasts with prior work that performs classification on full reports without localisation. Trained on over 2.7 million unlabelled chest X-ray reports and over 28,000 segmented and labelled reports, our model achieves state-of-the-art performance on report segmentation (0.0442 WinDiff) and multi-label classification (0.84 report-level macro F1) over 37 radiological labels and 8 NLP-specific labels. This work establishes new benchmarks for fine-grained understanding of free-text radiology reports, with precise localisation of semantics unlocking new opportunities to improve computer vision model training and clinical decision support. We open-source our annotation tool, model code and pretrained weights to encourage future research.

Keywords: Document Classification, Named Entity Recognition, Other (Multi-Task Learning, Multi-Label Classification, Chest X-rays, Radiological Report Segmentation)

1. Introduction

Chest radiography is one of the most commonly performed imaging examinations, with over 50 million chest X-rays annually in the United States alone. The interpretation of chest X-rays is critical for screening, diagnosis, and management of numerous pulmonary conditions as well as cardiovascular diseases (Annarumma et al., 2019; Ueda et al., 2023). However, manual review of chest X-rays can be an onerous process prone to errors due to the large volumes of images interpreted daily by radiologists, with reported diagnostic discrepancies in up to 30% of cases (Busby et al., 2018). This has motivated growing interest in using machine learning, such as deep learning with convolutional neural networks (CNNs), to automate chest X-ray interpretation and assist radiologists (Rajpurkar et al., 2017).

A major challenge in developing accurate deep learning models for chest X-ray analysis is insufficient and low quality labeled training data (Oakden-Rayner, 2020). While large archives of radiology images and reports exist in hospital PACS (Picture Archiving and Communication Systems) and public datasets such as MIMIC-CXR (Johnson et al., 2019), obtaining labels for the images is time-consuming and expensive, so NLP labels derived from free-text reports that accompany images are often used as proxy labels for training computer vision (CV) models. NLP labellers include earlier rule-based models such as NegBio (Peng et al., 2018) and CheXpert (Irvin et al., 2019), and recent transformer (Vaswani et al., 2017) models like CheXbert (Smit et al., 2020) built on top of the BERT (Devlin et al., 2019) model. However, the training labels for these transformer models are still largely extracted using CheXpert, thus replicating its limitations. The imperfect nature of the available NLP tools and labels has been noted in various works (Bressem et al., 2020a; Jain et al., 2021) and is restricting the capabilities of subsequent CV models trained on NLP labels.

In particular, current NLP methods are limited to classifying entire reports, lacking the localisation needed to link text observations precisely to image regions. This provides only weak supervision for CV and multi-modal model training, which is a limiting factor in applications such as the automatic image-to-text generation of chest X-ray reports (Jing et al., 2018; Qin and Song, 2022). To strengthen the localisation specificity and accuracy for medical imaging applications, there is a need for techniques that can jointly identify self-contained segments within a report and assign classification labels to each segment.

Furthermore, an important consequence of suboptimal NLP report classifiers is that generative models, such as image-to-text report generation models, cannot be adequately evaluated for their clinical accuracy. A recent review on chest X-ray report generation by Liu et al. (2023) showed that a significant portion of the proposed models were evaluated solely using natural language generation (NLG) metrics such as BLEU (Papineni et al., 2001). However, it is more important to evaluate the clinical accuracy of generated reports, making it difficult to measure progress in the field. This

LREC-COLING 2024, pages 862–875

is especially problematic given the recent proliferation of Large Language Models (LLMs) that can generate text with high fluency but are prone to hallucination, necessitating the development of a reliable model for evaluating clinical accuracy of the generated reports.

LLMs and multi-modal Foundational Models have also been applied to clinical and medical imaging data in research (Singhal et al., 2023; Wu et al., 2023; Tu et al., 2023), but they suffer from a few major weaknesses which we seek to address. Firstly, medical data is sensitive and unsuitable to be processed using proprietary APIs in practice, where (comparatively) lightweight and locally-hosted models may be preferred. Secondly, it is unclear what is included in the training corpus of such models, and they often forego the report classification task for report summarisation. As a result, it is unknown how they may perform on report classification or the proposed joint segmentation/classification task.

In this paper, we present a novel task of segmenting reports into semantically coherent segments, and propose a transformer-based approach for fine-grained analysis of free-text radiology reports which both identifies and classifies segments. Our aim is to segment a report into chunks that describe different radiological findings, and associate each chunk with an extensive set of pre-defined labels including all the most commonly reported findings. We make the following contributions:

- We present the new task of jointly segmenting radiology reports into semantically coherent chunks (segments) and assigning labels to each segment, with 37 radiological labels in our ontology.
- We present a novel transformer-based model tailored for radiology reports that can jointly extract segments from reports and perform multi-class multi-label classification on each segment.
- We perform a set of detailed experiments involving report segmentation and classification in single-task, pipeline and joint settings, showing the benefits of segmenting reports and jointly learning segmentation and classification.
- We make public our code and model weights from pretraining and fine-tuning to encourage further research and development. We hope that these resources will allow the research community to build on the fine-grained labelling of chest X-ray reports enabled by our work.

2. Related Work

Radiology Report Classification There has been growing interest in developing deep learning models for radiology data. For chest X-rays, Wang et al. (2017) developed an ontology of 8 common thoracic disease labels and released the ChestX-ray8 dataset containing 108,948 images with holistic disease multi-labels. Johnson et al. (2019) expanded this with the MIMIC-CXR dataset of 227,835 reports. In terms of NLP report classification models, early rule-based systems include NegBio (Peng et al., 2018) and CheXpert (Irvin et al., 2019), which identify 14 findings in radiology reports and classify them as positive, negative, uncertain or not mentioned. Neural models include bidirectional long short-term memory networks (BiLSTM) for report-level classification (Cornegruta et al., 2016). More recent transformer-based models like CheXbert (Smit et al., 2020) and CheXpert++ (McDermott et al., 2020) fine-tuned BERT on a combination of CheXpert and human labels to perform multi-label classification of chest X-ray reports. Bressem et al. (2020b) conducted both pretraining and fine-tuning for a BERT model on German language chest X-ray reports. RadBERT-CL (Jaiswal et al., 2021) performed pretraining using a contrastive loss objective and fine-tuning for multi-class multi-label classification. Cid et al. (2024) pretrained and fine-tuned a RoBERTa (Liu et al., 2019b) model for chest X-ray report classification on 45 labels.

Document Segmentation Various previous work have focused solely on segmenting documents texts into sections without assigning topic labels. For instance, Koshorek et al. (2018) utilised hierarchical BiLSTMs to formalise segmentation as a supervised task, Badjatiya et al. (2018) employed CNN and attention-enhanced BiLSTMs, and Wang et al. (2018) used restricted self-attention and conditional random field (CRF) (Lafferty et al., 2001) to perform discourse segmentation. More recently, Lukasik et al. (2020) proposed three BERT-based models for document and discourse segmentation.

Named Entity Recognition (NER) NER involves identifying textual mentions of predefined multi-token categories like people, locations and organizations. NER frequently used sequential models like LSTM-CRFs (Lample et al., 2016; Ma and Hovy, 2016), and more recent NER systems have adopted transformer networks like BERT, such as those proposed by Zhang et al. (2019). While NER focuses on identifying multi-token expressions, these are often fixed or formulaic. Our work performs segmentation of chest X-ray reports guided by radiological findings, rather than named entities. Our model must adapt to variable report-

ing styles rather than relying on consistent lexical triggers present in NER tasks.

We draw attention to both documentation and NER as the proposed segmentation task shares similarities to both, requiring almost complete coverage of entire reports while retaining the flexibility to exclude medically irrelevant information. Previous works in text segmentation generally modelled the task as a binary classification problem, where a positive label indicates that the current input (word/line/sentence) is the end of a segment, necessarily covering the entire document. In contrast, label schemes used for NER such as Beginning/Inside/Outside (BIO) offer the ability to more precisely extract salient segments pertaining to radiological findings. In addition, the structure inherently built in to the more detailed NER label schemes have been reported to improve performance, especially when the decoder is constrained to output legal transitions[1] (Gunawan et al., 2018; Lester et al., 2020). These properties make the NER label schemes more suitable for our task.

Joint Segmentation and Classification There have been some approaches that address jointly segmenting and classifying documents on general domain text. Tepper et al. (2012) outlined a feature-based approach using the BIO label scheme at a line level, testing both a joint (BIO-X) and a pipeline model for identifying sections in clinical records. Arnold et al. (2019) used BiLSTMs to predict the topic of sentences, then performs post hoc segmentation using the stepwise cosine difference of topic embeddings. Barrow et al. (2020) proposed the segment pooling network to dynamically generate segment embeddings for classification and corresponding ground-truth labels for training. They used scheduled teacher forcing (Williams and Zipser, 1989) with exploration to help the model converge, and jointly optimised the multi-task loss.

Joint modelling of segmentation with additional tasks has also been shown to improve performance compared to segmentation-only or pipeline approaches, suggesting that segmentation provides useful supervisory signal for other tasks. For example, Liu et al. (2021) performed joint segmentation and summarisation of news articles, and MT-DNN (Liu et al., 2019a) has been used for NER on biomedical data (Khan et al., 2020). Other works focus on document segmentation and used topic classification as an auxiliary task to improve segmentation performance, such as the use of hierarchical transformer for text segmentation(Lo et al., 2021) and multi-task transformer to section clinical notes (Zhang et al., 2022).

Concurrently to our work, related approaches have been proposed for SemEval 2023's Task 3.3 on multilingual detection of persuation techniques in online news (Piskorski et al., 2023), which is a paragraph-level multi-label classification task. While many methods simply used paragraph-level inputs to bypass the segmentation task, some relevant approaches include RoBERTa CRF (Pritzkau, 2023) and multi-head BERT (Baraniak and Sydow, 2023). We refer readers to Ojha et al. (2023) for a full list of entries.

In summary, to the best of our knowledge, previous works in radiology report classification operate on full reports rather than localising labels to precise segments informed by findings. In contrast, our work addresses extracting segments from chest X-ray reports and assigning radiological concepts from an expanded 37-label ontology to each segment. Through the addition of the segmentation task and subsequent classification at the segment rather than report-level, we aim to achieve higher classification performance and provide stronger localisation cues for medical imaging applications compared to existing report classification approaches.

3. Materials and Methods

3.1. Datasets and Radiological Ontology

Data Our chest X-ray report dataset contains 2,775,902 reports, of which 1,171,885 unique, across 259,152 patients collected from six UK hospitals between 2006-2019 in accordance with national governance procedures. The data covers a diverse population distribution of gender, age, number of images per patient, and time between images. Additional reports were obtained from MIMIC-CXR and Open-I (Demner-Fushman et al., 2012) datasets for pre-training, where non-empty sections were concatenated to create 1.53 million unique reports in total.

Radiology labels The reports were annotated at the segment level with a set of 45 labels, 37 of which correspond to a taxonomy[2] of findings for frontal chest x-rays. The taxonomy was developed in an iterative process involving discussions and annotations by a team of five radiologists, with a goal of identifying the smallest set of labels to capture key radiological findings. The remaining 8 labels capture report-specific information such as metadata, technical issues, recommendations and comparisons.

Report Annotation Three radiologists annotated a total of 28,521 chest radiograph reports, identifying non-overlapping segments and assigning la-

[1] For example, the start of a segment/entity must be tagged with the "beginning" label, so the transition OB is legal while OI is not.

[2] https://x-raydar.info/ontology

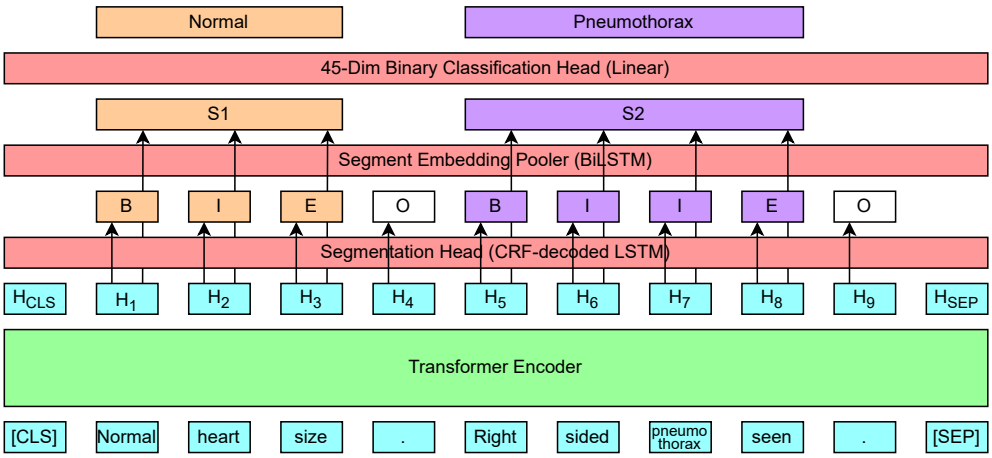

Figure 1: Proposed joint model architecture. A tokenised report is used as input into the transformer encoder. The resulting token embeddings are first used to predict the segment boundaries, based on which they are then pooled to form segment-level embeddings for classification. The losses from the segmentation and classification heads are weighted linearly.

bel(s) to each segment. This process extracted 127,809 segments, of which 78,465 were unique, resulting in an average of 4.48 segments per report. Example data is shown in Table 6, and additionally in Table 7 of the appendix.

To assess inter-annotator agreement, we evaluated a subset of 200 reports annotated by all three radiologists. For each report, consensus finding labels were derived by majority voting across the radiologists. Comparing the labels from each radiologist against the consensus yielded an average macro Matthew's correlation coefficient (MCC) of 0.8878. For segmentation, the mean pairwise WinDiff is 0.0614 (lower is better). Section 4 contains a further explanation of metrics used in this work. This high level of agreement indicates a reliable ground truth for developing and evaluating our model's ability to replicate human performance on segmenting and labeling chest X-ray reports. Additional information on the data and annotation process are available in Cid et al. (2024).

3.2. Task Description

Given a chest X-ray report, our task is to jointly extract text segments and their corresponding findings labels. Formally, for a report R consisting of n words/tokens $W = (w_1, ..., w_n)$, we predict for each word its segmentation label $\hat{Y} = (\hat{y}_1, ..., \hat{y}_n)$ where $\hat{y}_i \in \{I, O, B, E, S\}$[3]. $\hat{Y}$ is used to determine the boundaries of segments $\hat{S} = (\hat{s}_1, ..., \hat{s}_m)$, where each $\hat{s}_j$ is a contiguous subsequence of W. For example, if a report contains five words $(w_1, ..., w_5)$ with segmentation labels $(y_1, ..., y_5) =$ OBIES, then the corresponding segment boundaries would be (1, 4), (4, 5). Accordingly, $s_1 = (w_1, ..., w_3)$ and $s_2 = (w_4)$. For each predicted segment $\hat{s}_j$, we also predict its multilabel classification labels $\hat{z}_j \in \{0, 1\}^{45}$.

[3]This is using the IOBES scheme, where the labels correspond to Inside/Outside/Beginning/End/Singleton respectively.

3.3. Proposed Model Architecture

Our model architecture can be seen in Fig. 1. We use a transformer model, RoBERTa which has been pretrained on chest X-ray report data (appendix section B), to encode entire reports word by word. The output token embeddings are used for both segmentation and classification tasks in their respective task-specific heads. The segmentation head consists of one LSTM and one linear layer, with a CRF decoder enforcing the transition rules. The segmentation head outputs labels in the IOBES scheme, which are then converted into segment boundaries. The token embeddings are grouped into segments according to the obtained segment boundaries, and they are then pooled with a BiLSTM layer to produce the segment embeddings. The classification head is a single linear layer and outputs 45-dimensional binary segment-level finding labels.

3.4. Training Strategies

Teacher Forcing During the first epoch of training, we use teacher forcing on the segmentation labels if the model predicts an invalid segmentation. Since the classification task depends on the segment-level embeddings generated using the segmentation predictions, this allows the classification task to train when segmentation performance is poor. Typically, teacher forcing is used

with RNNs on text generation tasks, where each generated token is one timestep and teacher forcing is switched on and off at the token level. However, for our task we opt for teacher forcing to operate at the report level instead for two reasons. Firstly, our task involves only two timesteps (segmentation and classification), so the problem of compounding conditioning on erroneous predictions does not apply. Secondly, changing the segmentation prediction for random tokens has the potential to introduce illegal transitions which would result in segmentation errors. This could worsen downstream classification performance, running counter to the purpose of using teacher forcing.

Segment Label Alignment In order to generate the derived classification labels for incorrectly predicted segments for training and evaluation, we use a version of the greedy maximal overlap alignment algorithm (Barrow et al., 2020) modified to fit the task. Firstly, we allow predicted segments to map onto an empty segment (containing no labels) if it does not exist in the ground truth annotation. Secondly, we allow one-to-many and many-to-one alignments. This is to account for spans such as "Normal heart size. CTR XX/YY.", which may be labelled as one singular segment or as separate segments. We note that neither option is objectively correct or incorrect, which adds to the complexity of evaluating the segmentation task. The alignment process (and the segmentation predictions) are inherently noisy, possibly resulting in a slight reduction in classification performance.

Training Configurations and Hyperparameters are given in section C of the appendix.

4. Evaluation Metrics

Segmentation Task Consistent with existing literature on text segmentation, we report a sliding-window metric in WinDiff (Pevzner and Hearst, 2002). We set the window size to the recommended half of the average length of segments in the dataset, resulting in a window size of 6 words. Due to the calculations being done at a sliding-window-level, WinDiff assesses the segmentation performance locally. We note that WinDiff assumes a binary label scheme where segments cover the entire report and does not take into account "Outside" labels. This leads to an overestimation of segmentation performance on our dataset.

To address these issues, we additionally report segmentation boundary match rate (MR@B), which we calculate as the proportion of reports where the predicted segment boundaries exactly match the ground truth segmentation. This is an attempt to measure segmentation performance over entire reports, while also accounting for both starting and end points when evaluating segment boundaries.

Classification Task For classification, we report Matthew's correlation coefficient, precision, recall, F1 score, macro-averaged on both the segment and report levels. Report-level labels are generated for each report using the union of labels from its constituent segments for evaluation.

Joint Tasks: We use a multi-label variant of David Batista's entity-level NER evaluation[4], which is itself based on the MUC-5 error categories (Chinchor and Sundheim, 1993) and SemEval 2013 evaluation schema (Segura-Bedmar et al., 2013). The entity evaluation corresponds to the aforementioned segment-level multilabel classification evaluations. Notably, we define and use a version of the strict F1 measure to assess joint segmentation and classification performance. For strict F1, each predicted segment is counted as a true positive if and only if it matches the ground truth in both segment boundaries and all classification labels. A full description of the other proposed metrics is given in section A of the appendix.

5. Experiments

5.1. Baseline Report Classification

As a baseline, we test classification performance when entire reports are used as inputs. We also compare the performance of several publicly available transformer models pretrained on relevant domains: BERT-base, RoBERTa-base, CXRBert (Boecking et al., 2022), ClinicalBERT (Alsentzer et al., 2019), and PubMedBERT (Gu et al., 2022).

5.2. Segmentation Experiments

We examine whether off-the-shelf approaches are suitable for extracting segment boundaries. Splitting reports at full stops proved to be overly crude from preliminary testing, so we report performance from spaCy(Honnibal et al., 2020) Sentencizer which is a premade sentence splitter.

Effect of Pretraining First, we test the effect of in-domain pretraining on a baseline single-task segmentation model with a linear segmentation head and no decoder, using the BIO label scheme.

Segmentation Head and Label Scheme Furthermore, we explore the effect of the segmentation label scheme on the segmentation performance. For segmentation label schemes, we test BIO, IOBE and IOBES, where I-inside, O-outside, B-beginning, E-end, S-singleton. For segmentation head type, we test linear, LSTM and BiLSTM. For the decoder, we test CRF, constrained decoding

[4] https://github.com/davidsbatista/NER-Evaluation

(Lester et al., 2020), or no decoder. Additionally, transition legality can be enforced for the constrained and CRF decoders.

5.3. Segment Classification

Performance Gain over Report Classification: Similarly to the baseline report-level classification experiment, we test models where each input into the transformer encoder is one individual segment. This allows us to establish the expected performance gain from learning classification at the segment level. The aforementioned pretrained models are tested.

Embedding Pooler Type: The transformer model outputs an embedding for each token within the report. In order to compute the predicted classification labels at a segment level, we first use a pooling layer to aggregate the token embeddings into segment embeddings. We explore the effect of different pooling methods on downstream classification performance, where ground truth segmentation labels are used for each experiment. We test max, average, attention, LSTM, and BiLSTM pooling.

5.4. Joint Experiments

Pipeline Models We assess whether training for segmentation and classification jointly yields performance benefits over training two separate task-specific models. As a baseline, we test two pipeline setups, where the pipelines simply consist of the best performing single-task models from Sections 5.2 and 5.3. We use the segmentation model to produce the segmentation predictions, which can then be used to generate the predicted segment texts and corresponding aligned classification labels. For the classification component we test models trained for both single-segment input and report-input pooled-segment classification.

Teacher Forcing For joint models, losses from the segmentation task and classification task are initially weighted at 0.05:0.95. We firstly explore two teacher forcing strategies. By default, teacher forcing is only used for reports with no valid predicted segments, which can optionally be supplemented with a probabilistic schedule. Concretely, for each report in a training batch, the segment embeddings have probability p to be generated from ground truth segmentation labels and $1 - p$ from model predicted segmentation. The probability p decays linearly from 1 to 0 in the first epoch.

Task Weighting Finally, we perform a linear sweep over segmentation-classification task weightings between 0.05 and 0.95 to examine the tradeoff between the two tasks.

Model	WinDiff ↓	MR@B ↑
Spacy Sentencizer	0.0962	0.5704
roberta-base BIO linear	0.0513	0.7200
+ Pretraining	0.0490	0.7273
+ IOBES	**0.0463**	0.7294
+ LSTM	0.0478	0.7270
+ CRF	0.0497	0.7273
+ Enforce Transitions	0.0479	**0.7453**

Table 1: Single-task performance of segmentation models. Best results are indicated in **bold**.

6. Results and Discussion

Segmentation Task We show a reduced set of segmentation results in Table 1. Each additional row indicates a modelling decision and shows its effect on the segmentation performance. While WinDiff does not monotonously improve, the ability to enforce the legality of transitions with the use of a CRF decoder (final row) significantly improves the boundary match rate, which we believe to be better suited to our task as explained in Section 4. Our final segmentation setup was a CRF-decoded LSTM with enforced transitions using the IOBES label scheme.

Classification Task Table 2 shows the performance of baseline single-task classification models fine-tuned from different pretrained transformer model checkpoints. For all models tested, there is approximately a 0.05 increase in absolute report-level macro F1 when the model uses segment-level inputs. This is likely due to the more granular nature of the input texts and labels, which allow for more precise gradient updates.

Table 3 shows the effect of the pooling layer architecture on the final classification performance. We also include the best performing report and segment classification models (measured by validation metrics) from the previous section for comparison. For segment-level F1, most of the pooled segment classification models tested were able to improve upon the baseline single-segment classification model. Additionally, for report-level F1, most pooled models show better performance than the baseline model. This may suggest that segments within the same report contain complementary information, and the report-level input of the pooled models allow for such complementary information to be attended to across segments. It would then seem logical that the additional report-level context improves overall label coherence and results in higher performance at the report level compared to single-segment input models. At the segment level, the improvements could be explained by the fuzzy nature of the ground truth segmentation labels detailed in Section 3.4. Informa-

Input	Pretraining	F1-S	F1-R
report	bert-base	-	0.7793
	clinicalbert	-	0.7723
	cxrbert-general	-	0.5221
	pubmedbert	-	0.7905
	roberta-base	-	0.7766
	roberta + cxr	-	<u>0.7969</u>
segment	bert-base	0.8348	0.8420
	clinicalbert	0.8355	0.8437
	cxrbert-general	0.8373	0.8447
	pubmedbert	**0.8406**	**0.8482**
	roberta-base	0.8315	0.8394
	roberta + cxr	0.8343	0.8414

Table 2: Baseline single-task classification performance of several pretrained transformer models. Best results overall for each metric are indicated in **bold**, and the best report-level F1 achieved by a report-level classification model is indicated in <u>underline</u>.

tion contained in neighbouring segments may be especially useful when predicting segment-level classification labels. Overall, the pooled classification models do outperform the baseline models, despite some variations in performance between different pooling methods.

Examining the pipeline model results in Table 5 also shows that the pooled model (compared with the single-segment model) yields higher classification performance when predicted segments are used. The ability to attend to neighbouring segments likely improves resilience to segmentation errors. We move forward with a BiLSTM embedding pooler for classification because it achieved the best F1 and MCC at the segment level.

Joint Segmentation and Classification

Teacher Forcing Table 4 shows a reduced set of results from the teacher forcing experiment. Teacher forcing all samples while training (TF all), perhaps surprisingly, yielded the best results overall. Under this scheme, the segments are all extracted using the ground truth segmentation during training, but the loss from the segmentation task is still used to perform gradient updates. The higher report-level classification performance of TF all can likely be attributed to the benefit of consistently using ground-truth segmentations for the classification task. However, its lower segment-level classification performance may indicate a lower resilience to segmentation errors at test time.

Task Weighting Figure 2 shows the tradeoff between classification and segmentation performance under different weightings for the two losses. We test both TF invalid and TF invalid + decay stragies from the previous section. Segment-level macro F1 appears to be fairly stable around 0.81 but decreases rapidly once the segmentation loss is weighted above 80%. Boundary match rate shows a similar pattern and largely stabilises once segmentation loss weight is above 25%, with the TF-invalid strategy showing a slight downward trend. The non-monotonous change in these metrics in response to varying loss weightings, and especially their stability across a wide range of weightings, seems to support the idea that the two tasks contain complementary information and benefit from training jointly.

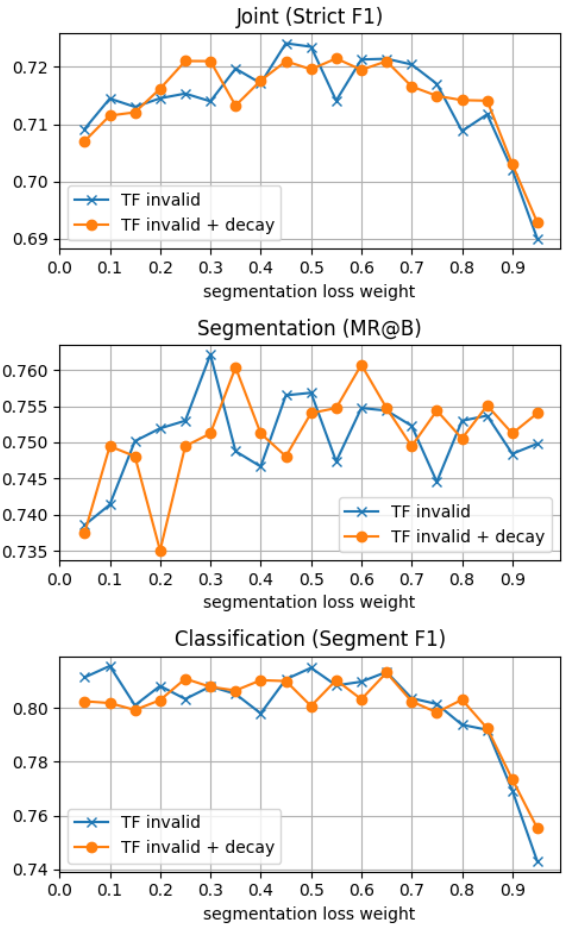

Figure 2: Tradeoff between segmentation and classification performance. The loss weighting of the segmentation task is varied between 0.05-0.95 in increments of 0.05.

Results Overview and Comparison Finally, in Table 5 we compare models introduced in this paper with published chest X-ray report classification results from previous works. However, other works use different taxonomies which makes direct comparison of metrics difficult. The single-task models provide very strong baselines, though the baseline segment classification models uses ground-truth segmentation labels at inference time, so their performance is likely unattainable in a real-

Modelling		Segment Level				Report Level			
Input	Pooling	Prec	Recall	F1	MCC	Prec	Recall	F1	MCC
Report	-	-	-	-	-	0.8178	0.7877	0.7969	0.7815
Segment	-	**0.8565**	0.8308	0.8406	0.8385	0.8599	0.8421	0.8482	0.8341
Report	Mean	0.8353	0.8137	0.8202	0.8184	0.8626	0.8235	0.8387	0.8255
	Mean Sqrt	0.8416	0.8336	0.8331	0.8314	0.8601	0.8427	0.8472	0.8336
	Max	0.8429	**0.8427**	0.8406	0.8378	0.8606	**0.8522**	**0.8544**	0.8396
	LSTM	0.8473	0.8354	0.8396	0.8366	0.8564	0.8444	0.8486	0.8336
	BiLSTM	0.8481	0.8392	**0.8412**	**0.8386**	0.8582	0.8484	0.8509	0.8364
	Attention	0.8423	0.8414	0.8382	0.8360	**0.8636**	0.8513	0.8540	**0.8399**

Table 3: Performance of single-task classification models with different token embedding to segment embedding pooling methods. Ground truth segmentation is used. We also report expanded baseline performance of the best models from table 2 for comparison. Best results are indicated in **bold**.

Strategy	F1-S	F1-R	MR@B	F1-Str
TF all	0.8119	**0.8412**	**0.7467**	**0.7163**
TF invalid	0.8114	0.8383	0.7386	0.7092
+ decay	**0.8132**	0.8395	0.7442	0.7125

Table 4: Reduced teacher forcing results from three notable strategies. F1-Str is the joint strict F1 metric. Best results are indicated in **bold**.

world setting. Nevertheless, the pipeline segment classification models, which use predicted segmentations as input, are already able to improve report-level F1 by 2-4% absolute compared to performing classification at the report level. Finally, our jointly trained model was able to improve upon the single-task and pipeline models in terms of both segmentation and classification performance, and approaches the theoretical maximum report-level performance set by the segment classification model. The gain from jointly training for segmentation and classification likely suggests that the tasks contain complementary information, which is in line with findings from previous works. In this work, the segmentation and subsequent pooling focuses the classification signal, allowing for more precise weight updates. Conversely, learning to classify each segment makes the model more sensitive towards topic changes between contiguous segments, improving the segmentation performance.

Error Analysis For the segmentation task, a common error is when a sentence is predicted as two segments while being labelled as a single segment in the ground truth and vice versa (subset and superset error types, respectively), such as sentences in the form of "X and Y" where X and Y are radiological findings. An example of this can be seen in Table 6. Neither interpretation is necessarily objectively incorrect in these cases, though extracting these as two individual segments is more in line with our objectives and downstream applications. We observe an inverse correlation between these two error types as expected, recording 482 subset and 544 superset errors out of 13,082 ground truth segments in the test set.

Similarly for classification, we generally observe lower performance in labels which are not precisely grounded in imaging features, and may therefore be prone to the annotators' differing interpretations, commonly the NLP-specific labels. An example of this is the "abnormal non clinically important" label, attaining 0.755 report-level F1 which is lower than the 0.840 macro average.

Computing the confusion matrix for 45-label multi-label multi-class classification is challenging as it would result in 2^{45} possible combinations. However, we can isolate smaller groups of similar labels which are often confused with each other, defined as segments where label X is a false positive and label Y is a false negative. For example, volume loss and atelectasis, where 28/375 cases (7.5%) and 24/317 cases (7.6%) are confused for the other label. These may arise from from differences in labelling due to difference in age, training, or hospital convention between the radiologists. We also observe low performance for some extremely rare labels. In the test set with 2,846 reports, paraspinal mass was positive in 5 reports (0.18%) with a segment-level F1 of 0.44, and cardiac calcification was present in 13 reports (0.46%) with a segment-level F1 of 0.47.

7. Conclusion & Future work

In this work, we set out to improve the granularity and performance of chest X-ray report classification models due to the importance of their outputs on research in this field. For this purpose we have defined the task of joint segmentation and classification of chest X-ray reports, and have

Model Category	Model Name	Segmentation WinDiff	Segmentation MR@B	Classification F1-S	Classification F1-R	Classification MCC-R	Joint F1-Strict
Previous Works	CheXbert (A)	-	-	-	0.798	-	-
	GER-BERT (A)	-	-	-	0.92	0.89	-
	RadBERT-CL (A)	-	-	-	0.804	-	-
	X-Raydar-NLP (B)	-	-	-	0.845	0.841	-
Annotator Baseline	Mean agreement	0.0614	0.658	-	-	0.888	0.643
Single Task Best	Segment Classification (S)	-	-	0.841	0.848	0.834	-
	Segment Classification (P)	-	-	0.841	0.851	0.836	-
	Report Classification	-	-	-	0.797	0.782	-
	Segmentation	0.0479	0.745	-	-	-	-
Pipeline	Single Segment	0.0479	0.745	0.797	0.821	0.805	**0.724**
	Pooled Segment	0.0479	0.745	**0.810**	0.835	0.819	0.717
Proposed	Joint	**0.0442**	**0.755**	**0.810**	**0.840**	**0.824**	0.722

Table 5: Overview of performance of models introduced in this paper, compared with some baseline results. A: different taxonomy. B: different test set. S (single-segment input) and P (pooled report input): use ground truth segmentation for inference, so performance is likely unattainable in a real setting. Best results from comparable are indicated in **bold**.

Labelled	Bilateral widespread reticular shadowing and small volume lungs in keeping with ILD. (interstitial shadowing, possible diagnosis, volume loss) Dilated tortuous oesophagus as seen previously (abnormal non clinically important)
Predicted	Bilateral widespread reticular shadowing (interstitial shadowing) and small volume lungs (volume loss) in keeping with ILD. (possible diagnosis) Dilated tortuous oesophagus as seen previously (widened mediastinum)

Table 6: Example of annotations and predicted outputs for a single report. The first sentence was annotated as one segment with three labels, whereas the model predicts it as three contiguous segments each with their own label. The second sentence/segment was labelled "abnormal non clinically important" as the abnormality is unchanged, whereas the model predicts its original label instead.

proposed a multi-task transformer-based model to perform this task. We conducted extensive experiments to justify architectural decisions, and compared our proposed model with single-task baselines and state-of-the-art models. We demonstrate the benefit from the addition of the segmentation task, i.e. that it allows the classification task to be conducted with higher granularity, resulting in higher performance. We believe the benefit of the segmentation task to be generalisable to other taxonomies and datasets.

For future work, we suggest an exploration of additional auxiliary tasks such as token and report classification, for which the labels can be generated trivially. Furthermore, a more sophisticated multi-task learning loss weighting mechanism may help steer the gradient in a direction which is beneficial to both/all tasks. For segmentation, a robust aggregation technique and a method to determine consensus may help improve data quality.

More broadly, the segmented and labelled reports can enable mapping of individual text mentions to precise image regions annotated with clinical concepts like anatomical locations and pathological findings, facilitating precise multi-modal alignment. Our model can also be used to evaluate the clinical accuracy of chest X-ray report generation/summarisation models, with improved granularity and label coverage compared to previous works. In clinical settings, a reliable granular classification model can be used for second/double reading of reports, which can provide real-time feedback for the reporting radiologist as to how the report may be interpreted. This is particularly helpful in a learning situation, where resident radiologists' reports and findings are checked by a more experienced colleague.

Code Availability

The code and model weights will be made available on GitHub[5]. By releasing open-source code

[5]https://github.com/yznlp/RoBERTaX

with pretrained and fine-tuned weights for our models, we aim to facilitate further research and development in this field, allowing researchers and developers to build upon our work and create more advanced solutions for processing radiological reports. Unfortunately, we are unable to make the dataset publically available at this point due to the agreement with our data providers.

Ethics and Limitations

This study is based on non-identifiable data from three NHS Trusts (Hospital Networks) in the UK (6 hospitals) and the publicly available datasets MIMIC-CXR and Open-I. Data obtained from the three NHS Trusts followed national governance (GAfREC) and NHS data opt out procedures; these data were previously collected, non-identifiable information from patients as part of their care and support. The need for ethical approval and participant consent was waived since all of the data were anonymised. University Hospitals Coventry and Warwickshire NHS Trust was the primary NHS sponsor for this study (GAfREC ID: GF0274). The MIMIC-CXR public dataset was originally approved by the Institutional Review Board of the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA and used here under the PhysioNet Credentialed Health Data Use data use agreement.

The main limitation of this work is the challenge in choosing a suitable joint segmentation and classification metrics. While the proposed strict F1 measure was able to capture the performance of both tasks simultaneously and did behave as expected as shown in Figure 2, its robustness remains to be seen. In particular, requiring exact match on the segment boundaries and the full set of classification labels may encourage overfitting. This is especially important considering the templated nature of chest X-ray report data, particularly when modelled at the segment level. We are able to see the effect of this in Table 5, where the final model achieves a lower strict F1 despite outperforming the pipeline models in the segmentation and classification tasks individually. As a result, it was difficult to select a final model.

Bibliographical References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings. *arXiv:1904.03323 [cs]*. ArXiv: 1904.03323.

Mauro Annarumma, Samuel J. Withey, Robert J. Bakewell, Emanuele Pesce, Vicky Goh, and Giovanni Montana. 2019. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology*, 291(1):196–202. Publisher: Radiological Society of North America.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Pinkesh Badjatiya, Litton J. Kurisinkel, Manish Gupta, and Vasudeva Varma. 2018. Attention-based Neural Text Segmentation. *arXiv:1808.09935 [cs, stat]*. ArXiv: 1808.09935.

Katarzyna Baraniak and M Sydow. 2023. Kb at SemEval-2023 task 3: On multitask hierarchical BERT base neural network for multi-label persuasion techniques detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1395–1400, Toronto, Canada. Association for Computational Linguistics.

Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas Oard, and Philip Resnik. 2020. A Joint Model for Document Segmentation and Segment Labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 313–322, Online. Association for Computational Linguistics.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. volume 13696, pages 1–21. ArXiv:2204.09817 [cs].

Keno K. Bressem, Lisa C. Adams, Christoph Erxleben, Bernd Hamm, Stefan M. Niehues, and Janis L. Vahldiek. 2020a. Comparing different deep learning architectures for classification of chest radiographs. *Scientific Reports*, 10(1):13590. Number: 1 Publisher: Nature Publishing Group.

Keno K Bressem, Lisa C Adams, Robert A Gaudin, Daniel Tröltzsch, Bernd Hamm, Marcus R Makowski, Chan-Yong Schüle, Janis L Vahldiek, and Stefan M Niehues. 2020b. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics*, 36(21):5255–5261.

Lindsay P. Busby, Jesse Courtier, and Christine M. Glastonbury. 2018. Bias in radiology: The how and why of misses and misinterpretations. *Radiographics : a review publication of the Radiological Society of North America, Inc*.

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Yashin Dicente Cid, Matthew Macpherson, Louise Gervais-Andre, Yuanyi Zhu, Giuseppe Franco, Ruggiero Santeramo, Chee Lim, Ian Selby, Keerthini Muthuswamy, Ashik Amlani, Heath Hopewell, Das Indrajeet, Maria Liakata, Charles E. Hutchinson, Vicky Goh, and Giovanni Montana. 2024. Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. *The Lancet. Digital Health*, 6(1):e44–e57.

Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. 2016. Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks. *arXiv:1609.08409 [cs, stat]*. ArXiv: 1609.08409.

Dina Demner-Fushman, S. Antani, Matthew S. Simpson, and G. Thoma. 2012. Design and Development of a Multimodal Biomedical Information Retrieval System. *J. Comput. Sci. Eng.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23. ArXiv:2007.15779 [cs].

W Gunawan, Derwin Suhartono, Fredy Purnomo, and Andrew Ongko. 2018. Named-entity recognition for indonesian language using bidirectional lstm-cnns. *Procedia Computer Science*, 135:425–432.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv:1901.07031 [cs, eess]*. ArXiv: 1901.07031.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A. Young, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. 2021. VisualCheXbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 105–115. Association for Computing Machinery, New York, NY, USA.

Ajay Jaiswal, Liyan Tang, Meheli Ghosh, Justin Rousseau, Yifan Peng, and Ying Ding. 2021. Radbert-cl: Factually-aware contrastive learning for radiology report classification.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the Automatic Generation of Medical Imaging Reports. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586. ArXiv: 1711.08195.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Muhammad Raza Khan, Morteza Ziyadi, and Mohamed AbdelHady. 2020. MT-BioNER: Multi-task Learning for Biomedical Named Entity Recognition using Deep Bidirectional Transformers. *arXiv:2001.08904 [cs, stat]*. ArXiv: 2001.08904.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text Segmentation as a Supervised Learning Task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Brian Lester, Myle Ott, Luke Zettlemoyer, and Paul Michel. 2020. Self-training sampling with monolingual data uncertainty for neural machine translation. In *Proceedings of EMNLP*, pages 4410–4421.

Chang Liu, Yuanhe Tian, and Yan Song. 2023. A Systematic Review of Deep Learning-based Research on Radiology Report Generation.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-Task Deep Neural Networks for Natural Language Understanding. *arXiv:1901.11504 [cs]*. ArXiv: 1901.11504.

Yang Liu, Chenguang Zhu, and Michael Zeng. 2021. End-to-End Segmentation-based News Summarization. *arXiv:2110.07850 [cs]*. ArXiv: 2110.07850.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray L. Buntine. 2021. Transformer over Pre-trained Transformer for Neural Text Segmentation with Enhanced Topic Coherence. In *EMNLP*.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text Segmentation by Cross Segment Attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716, Online. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Matthew B. A. McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. 2020. CheXpert++: Approximating the CheXpert labeler for Speed,Differentiability, and Probabilistic Output. *arXiv:2006.15229 [cs, stat]*. ArXiv: 2006.15229.

Luke Oakden-Rayner. 2020. Exploring large-scale public medical image datasets. *Academic Radiology*, 27(1):105–112.

Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors. 2023. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, Toronto, Canada.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald M Summers, and Zhiyong Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188.

Lev Pevzner and Marti A. Hearst. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28(1):19–36. Place: Cambridge, MA Publisher: MIT Press.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Albert Pritzkau. 2023. NL4IA at SemEval-2023 task 3: A comparison of sequence classification and token classification to detect persuasive techniques. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 794–799, Toronto, Canada. Association for Computational Linguistics.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, Dublin, Ireland. Association for Computational Linguistics.

P Rajpurkar, J Irvin, K Zhu, B Yang, H Mehta, T Duan, D Ding, A Bagul, C Langlotz, K Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. 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)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. Publisher: Nature Publishing Group.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. *arXiv:2004.09167 [cs]*. ArXiv: 2004.09167.

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012. Statistical Section Segmentation in Free-Text Clinical Records. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2001–2008, Istanbul, Turkey. European Language Resources Association (ELRA).

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Generalist Biomedical AI. ArXiv:2307.14334 [cs].

D Ueda, T Matsumoto, S Ehara, et al. 2023. Artificial intelligence-based model to classify cardiac functions from chest radiographs: a multi-institutional, retrospective model development and validation study. *Lancet Digital Health*, 5(8).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471. ArXiv: 1705.02315.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward Fast and Accurate Neural Discourse Segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data. ArXiv:2308.02463 [cs].

Fan Zhang, Itay Laish, Ayelet Benjamini, and Amir Feder. 2022. Section Classification in Clinical Notes with Multi-task Transformers. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 54–59, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xiaohui Zhang, Yaoyun Zhang, Qin Zhang, Yuankai Ren, Tinglin Qiu, Jianhui Ma, and

Qiang Sun. 2019. Extracting comprehensive clinical information for breast cancer using deep learning methods. *International Journal of Medical Informatics*, 132:103985.