AI

MDPI

Article

Effective Machine Learning Techniques for Non-English Radiology Report Classification: A Danish Case Study

Alice Schiavone [1,*,†], Lea Marie Pehrson [1,2,3,†], Silvia Ingala [2,4], Rasmus Bonnevie [5], Marco Fraccaro [5], Dana Li [2,3], Michael Bachmann Nielsen [1,2,3] and Desmond Elliott [1]

[1] Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark
[2] Department of Diagnostic Radiology, Copenhagen University Hospital Rigshospitalet, 2100 Copenhagen, Denmark
[3] Department of Clinical Medicine, University of Copenhagen, 2100 Copenhagen, Denmark
[4] Cerebriu A/S, 1434 Copenhagen, Denmark
[5] Unumed Aps, 1055 Copenhagen, Denmark
* Correspondence: alsc@di.ku.dk
† These authors contributed equally to this work.

Abstract: Background: Machine learning methods for clinical assistance require a large number of annotations from trained experts to achieve optimal performance. Previous work in natural language processing has shown that it is possible to automatically extract annotations from the free-text reports associated with chest X-rays. Methods: This study investigated techniques to extract 49 labels in a hierarchical tree structure from chest X-ray reports written in Danish. The labels were extracted from approximately 550,000 reports by performing multi-class, multi-label classification using a method based on pattern-matching rules, a classic approach in the literature for solving this task. The performance of this method was compared to that of open-source large language models that were pre-trained on Danish data and fine-tuned for classification. Results: Methods developed for English were also applicable to Danish and achieved similar performance (a weighted F1 score of 0.778 on 49 findings). A small set of expert annotations was sufficient to achieve competitive results, even with an unbalanced dataset. Conclusions: Natural language processing techniques provide a promising alternative to human expert annotation when annotations of chest X-ray reports are needed. Large language models can outperform traditional pattern-matching methods.

Keywords: AI for healthcare; natural language processing; radiology report classification

Academic Editor: Miguel Molina-Solana

Received: 20 December 2024
Revised: 27 January 2025
Accepted: 5 February 2025
Published: 17 February 2025

Citation: Schiavone, A.; Pehrson, L.M.; Ingala, S.; Bonnevie, R.; Fraccaro, M.; Li, D.; Nielsen, M.B.; Elliott, D. Effective Machine Learning Techniques for Non-English Radiology Report Classification: A Danish Case Study. *AI* **2025**, *6*, 37. https://doi.org/10.3390/ai6020037

1. Introduction

Artificial intelligence research has long been working towards the automation of medical image interpretation with high accuracy and efficiency. Supervised machine learning algorithms show promising results in terms of diagnostic accuracy, while taking little time to identify a pre-defined set of abnormalities [1–3]. These algorithms are data-hungry, meaning that they need to be *fed* thousands of images to perform in the best way possible. Each case needs an *annotation*, or set of labels, that indicates whether abnormalities are present or not. These labels are expensive to obtain, as only human experts can become annotators. Concerns arise when a clinician's workload shifts from diagnostics to annotation, creating a new problem rather than solving it [4,5]. However, some information is already available when images are obtained from clinics, as radiologists describe their findings in a detailed free-text report.

Natural language processing (NLP) techniques have proven to be effective in automatically detecting the findings from a radiology report [6,7]. The use of these methods can aid us in acquiring the annotations required for the development of computer vision systems. Examples of possible techniques include string-matching rules which identify pre-defined patterns to output annotations, as well as state-of-the-art deep learning methods.

Rule-based systems use *regular expressions* (RegExs), a set of textual matching rules, to match patterns in phrases which indicate the presence or absence of a clinical finding, in addition to the type of presence, i.e., a positive or negative mention [8–10]. Rule-based systems require carefully crafted RegEx rules that are tailored to the target domain and language. Consider the sentence "`We exclude the presence of pneumothorax`", which contains a negatively mentioned finding. A simple RegEx like `(s/pneumothorax)` would match the word "pneumothorax" and wrongly interpret it as a positive mention. Modeling negations and spelling errors that naturally occur in free-text language is the main challenge of using RegEx annotation methods.

NegBio [11] is a RegEx method that detects negations and uncertainty using special language patterns. It inspired the methods used to extract labels from reports of chest X-ray images in English [12,13] and other languages such as Brazilian Portuguese [14], Vietnamese [15], and German [16]. *Large language models* (LLMs) have shown strong results in previous work on this task: *CheXbert* [17] is a BERT-based [18] language model trained on labels extracted using another rule-based method and human expert annotations, achieving better performance than the previous state-of-the-art rule-based methods.

Most of the available automatic annotation tools work only on radiology reports written in English. When working in other languages, fewer technical resources are available, which necessitates the development of task- and language-specific methods. The severity of this challenge can be quantified with reference to the European Language Equality Programme, which defines the *technological DLE factor* to measure the level of technological support for a language in terms of the available data and tools [19]. Although Danish is *not* considered a low-resource language, its technological DLE factor (according to the European Language Grid, release 3) is only 10,439, in contrast to 78,335 for English. It can also be compared to German and Spanish, for which automatic annotation tools have been developed, which have DLE factors of 39,929 and 36,751, respectively. The lack of resources for languages such as Danish makes the development of an annotation tool for radiology reports more difficult.

This study explored techniques for the efficient and cost-effective automatic annotation of chest X-rays reports written in Danish and proposes a machine learning model training strategy that matches the performance of similar English methods. With this in mind, the aim of this work is the following:

(1) Develop a string-matching algorithm for the automatic identification of positive and negative mentions of abnormalities in Danish chest X-ray reports.
(2) Compare the performance of the above method against various BERT-based machine learning models, including models trained on multi-lingual datasets versus models fine-tuned on the target language.
(3) Present an overview of the experimental results on the annotation effort needed to achieve the desired performance.

2. Materials and Methods

Ethical approval was obtained on 11 May 2022 from the Regional Council for Region Hovedstaden (R-22017450). Approval for data retrieval and storage was obtained on 19 May 2022 from the Knowledge Center on Data Protection Compliance (P-2022-231).

2.1. Materials: Data Collection

The data for this study were a subset of a larger dataset extracted from 2010–2019 from the PACS system of a major network of hospitals and private clinics in Denmark ($n = 11$). The dataset contained 1,452,561 unique cases and 878,305 associated reports, representing 808,662 studies with 357,205 unique patients. From this, 547,758 unique text reports were selected for annotation, of which a random subset of 2475 entries were selected for human expert labeling by medical personnel. See Figure 1 for an overview of the annotation process. The site with the most reports contributed ca. 139,000 examples (25% of the total). The site with the fewest reports contributed ca. 10,000 examples (2% of the total).

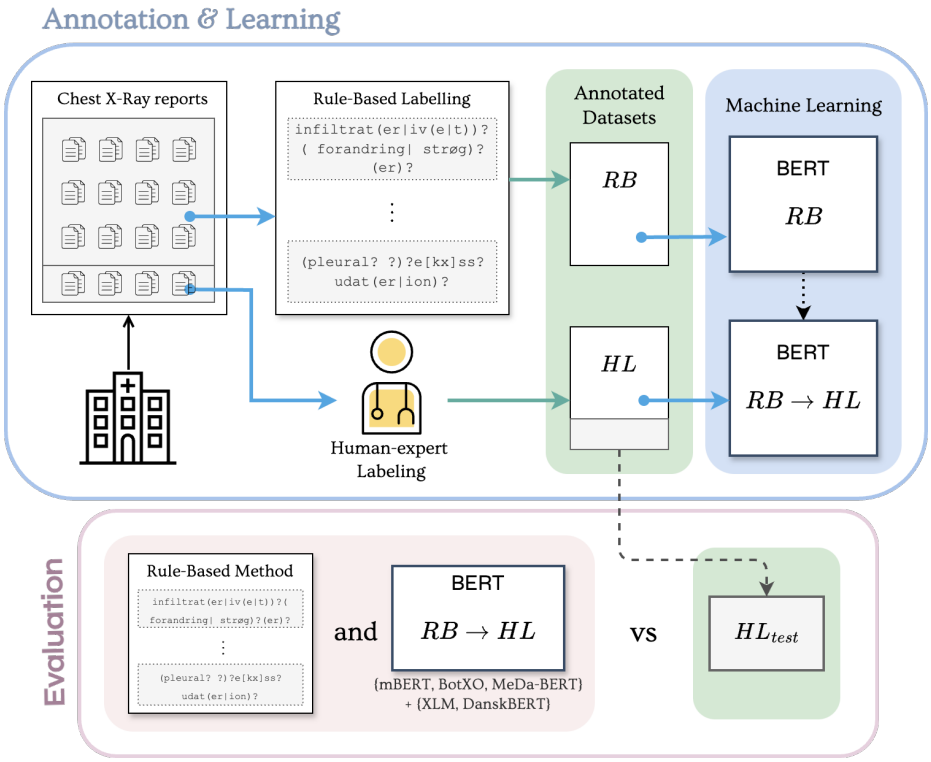

Figure 1. A total of 547,758 Danish chest X-ray reports were collected from a major hospital network in Denmark, covering 11 sites. The reports were subdivided into two sets: the first was annotated by *RegEx* rules (*RE*) and formed the *RB* (rule-based labels) dataset; the second and smaller set was manually labeled by expert human annotators, forming the *HL* set (human expert labels). The *RB* set was used to fine-tune a *BERT*-like model to annotate reports as one of forty-nine findings, as either a positive or negative mention or as a finding not mentioned. This model was then fine-tuned on the *HL* set. The *RegEx* and *BERT* models were then evaluated against a subset of the *HL* set that was not seen during training.

The label hierarchy used in our dataset was developed by Danish radiologists to match their diagnostic terminology [20–22]. It includes parent labels and increasingly specific sub-labels, with definitions to ensure consistent application. The hierarchy was designed to cover all findings comprehensively and to minimize the application of a catch-all *Other* label. The hierarchy was iteratively refined with an external medical software company through testing and comparisons with public datasets. Radiologists made all final label decisions, drawing on related studies [12,23]. Listing A1 in Appendix A presents the full hierarchy of labels.

In the report labeling process, each abnormality (finding) could be assigned to one of the following classes: a "positively mentioned finding" (a finding is present), "negatively mentioned finding" (no evidence of a finding is present), or "the finding not mentioned in the report" (the models were trained to identify this class but not evaluated against it).

Positive mentions were sometimes found with contextual modifiers in the reports, such as annotators noting uncertainty about the classification or using adjectives or comments to grade the severity of the finding. Since these mentions were still positive in the sense of describing the presence of a finding, they were coded as positive.

2.2. Methods: Regular Expressions (RE)

Inspired by the work by Irvin et al. [12], 574,758 reports in our dataset were labeled using a simple pattern-matching method-based NegEx [24], which we refer to as *RE* (Figure 2). Following this protocol, a total of 360 RegEx rules were created, which unfolded on 1,548,505 unique raw string patterns.

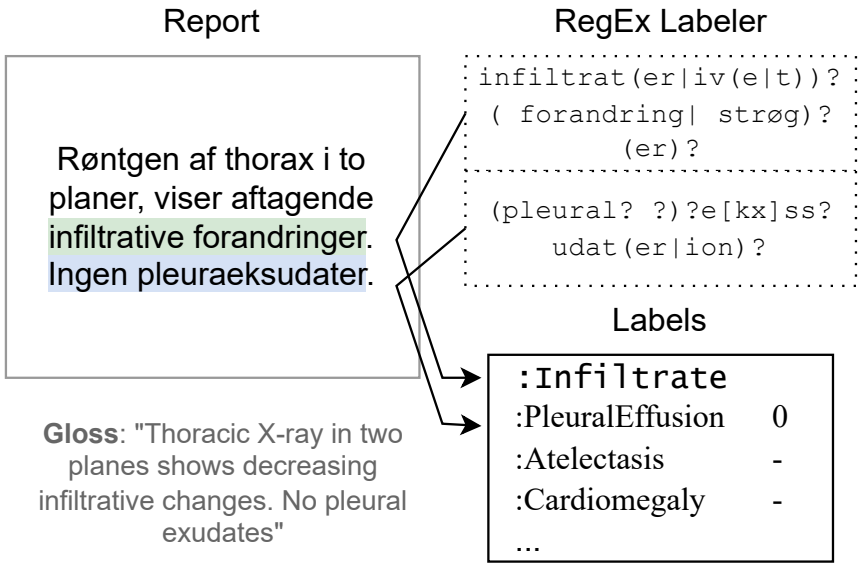

Figure 2. Constructed example of two RegEx rules matching a Danish chest X-ray report. The Regex Labeler outputs a set of findings as being positively (1) or negatively (0) mentioned or not mentioned (-). In green, the `Infiltrate` rule matches a positive mention. In blue, the `PleuralEffusion` finding is mentioned and negated by the word "ingen". Abnormalities that are not matched by any rule are assigned the "not mentioned" class.

2.3. Methods: Pre-Trained Language Models

Large language models (LLMs) are the current state of the art in many NLP tasks [25]. The training strategy developed by Smit et al. [17] to classify chest X-ray reports in English using BERT-like models was followed to classify Danish reports. Five LLMs pre-trained on Danish text were selected. The order of the presentation of the models' results is based on the amount of Danish text used in the model's pre-training, from the least to the most (Table 1). Box 1 summarizes terminology related to LLMs used in this paper.

Box 1. Large language models (LLMs) terminology.

Pre-training	Models are trained with unstructured, unlabeled data by predicting missing parts of the input as a next word prediction task.
Fine-tuning	A pre-trained LLM is further trained to specialize it for a specific task, e.g., radiology report classification.
Token	Fundamental unit of text in LLMs, representing words or subwords.
Tokenizer	A function that converts the input text into tokens.
Context	Amount of text, in tokens, that the model can process at any one time.

Table 1. Summary of the size (as in the number of parameters) and data used to pre-train the selected BERT-like models. 🌍 indicates multi-lingual data, 🇩🇰 stands for Danish, and ❄️ for Danish medical data. The models are presented from the least amount of Danish data used in their pre-training (*mBERT*) to the most (*DanskBERT*).

Pre-Training Data	Model Name	Parameters	Tokens/Words
🌍	mBERT	110 M	≈200 M words [1]
🇩🇰	BotXO	110 M	1.6 B words
🇩🇰❄️	MeDa-BERT	110 M	+123 M tokens
🌍	XML-RoBERTa	125 M	7.8 B tokens
🇩🇰	DanskBERT	125 M	+1 B words

[1] This number is a rough estimate, as the specific number was not reported by [18].

Since the primary focus of this study was to explore the feasibility of translating research techniques developed for English to a non-English language (Danish), LLMs with billions of parameters were excluded. This practical approach ensured that the findings in this paper are relevant and applicable for organizations with similar data, hardware, human resources, and privacy limitations. More details about pre-trained language models and the RegEx-based method are available in Appendix B.

2.4. Experimental Protocol

Using *RE* rules, labels were extracted from 547,758 text reports and collected in the *rule-based RB* dataset. This set was divided with a 98/2 split for validation during training.

The *human expert labeled* (*HL*) dataset with 2475 samples was divided into `train` ($n = 1600$), `dev` ($n = 125$), and `test` ($n = 750$) sets. The splits were generated using stratification for multi-label data as described by Sechidis et al. [26] and implemented using the `iterative-stratification` Python library (version 0.1.7). *RB* labels were generated in HL_{test} to compare *RE* against machine learning models. The datasets presented a huge imbalance between the findings and assigned classes (Figures 3 and A1).

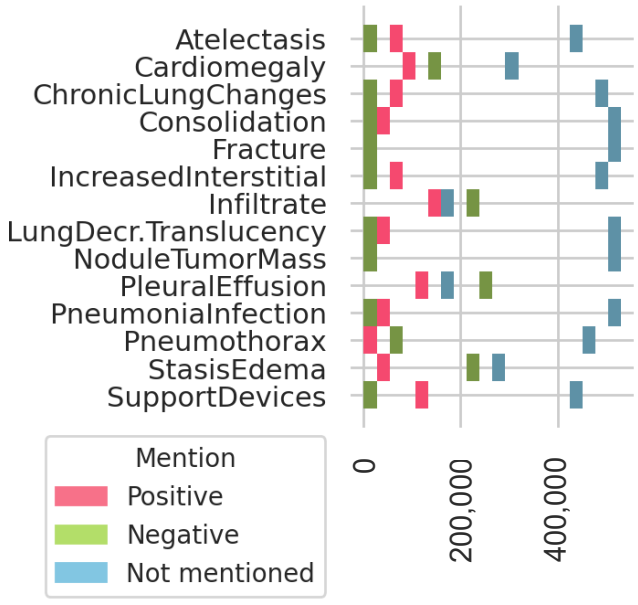

Figure 3. Distribution of the number of reports by the class assigned to each finding in the combined *RB* and *HL* datasets for the *most frequent findings*. For most abnormalities, "not mentioned" was the most frequent class, except for *Infiltrate* and *PleuralEffusion*, for which negated mentions were more common.

All language model weights were initialized in the respective pre-trained architecture, which produced a CLS token. This token was the input to a multi-label, multi-class linear prediction layer. Following the labeling protocol, this translated into a layer of 49 linear heads that could predict three different classes: a positive mention, negative mention, and no mention. The input text was tokenized, and only the first 512 tokens were used as the context for making predictions. In the dataset, a very small number of reports exceeded this limit, but the exact number depended on the model's tokenizer. All models were trained with unfrozen weights, using the cross-entropy loss and AdamW optimization with an initial learning rate of 1×10^{-5}. The loss was computed by summing the individual losses of the linear heads. All models were trained using one NVIDIA RTX 4090 GPU, with a batch size of 28 for RoBERTa base models and 30 for BERT base models.

2.4.1. Rule-Based Labels (RB)

A given machine learning model was trained on RB_{train}. The models' performance was tracked on RB_{dev}. All models were trained for 8 epochs, with an average wall clock training time of 6 h.

2.4.2. Human Expert Labels (HL)

A given machine learning model was trained on a subset of the data that was labeled by human experts (HL). HL_{train} was much smaller than RB; therefore, the models were trained for 50 epochs to improve the convergence. HL_{dev} was used for validation, on which the models showed performance degradation when training for more epochs.

2.4.3. Transfer Learning ($RB \rightarrow HL$)

Starting with models trained on the RB set, the models were further trained on the human expert-annotated data in HL, following the same hyper-parameters as the models trained exclusively on HL. This experiment is indicated in the results as $RB \rightarrow HL$. This technique is also called *transfer learning* (Figure 4), as in when a model trained on one task is then adapted to the target task by training further on a limited dataset from the desired domain. The main results report the average performance over 5 repeated training runs with different random seeds to address training variation.

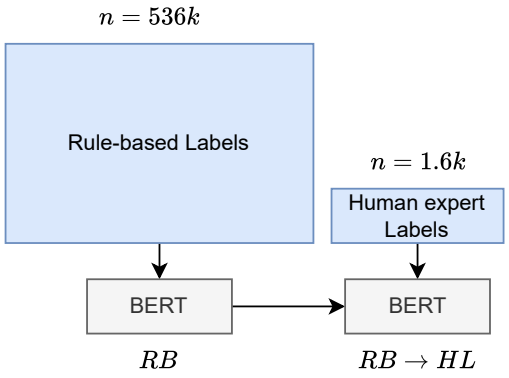

Figure 4. A large-scale dataset annotated with rule-based labels (RB) was used to tune a BERT-like model to predict 49 findings in Danish chest X-ray reports. This model was then fine-tuned on a smaller set of different reports labeled by human expert annotators (HL).

2.5. Resource-Driven Experiments

This study examined training strategies that could enhance performance in scenarios with limited manual labeling or data resources. Additionally, the robustness of the methods was evaluated through the use of model ensembles.

2.5.1. Definition of Most Frequent Findings

A subset of the initial findings was selected so that it corresponded to the top ten most frequent annotations for the positive and negative classes in the entire dataset, as annotated by human expert annotators and using the *RE* method. This subset resulted in 14 findings and was defined only for evaluations, meaning that no model was trained exclusively on this subset. Comparing the performance of only the most frequent labels allowed us to compare the Danish models to English methods that use smaller label sets.

2.5.2. Definition of Model Ensemble

Model ensembles are a well-established machine learning method for increasing the accuracy of predictions by training different models and then averaging their predictions [27]. They have been shown to outperform single classifiers within the ensemble, as they tend to decrease the generalization error without increasing the model variance [28]. In the second stage of the language model training, a model ensemble was trained using 5-fold cross-validation. HL_{train} was divided into 5 folds of 320 examples each: 4 folds were used to train one model, and the left-out fold was used for validation. This process was repeated to train 5 models on their respective training–validation folds. For testing, the predictions on HL_{test} from each model were averaged using majority voting.

2.5.3. Definition of Data Ablation

Knowing how many expert annotations are needed to develop a good annotation tool is crucial. The $RB \rightarrow HL$ experiment was repeated by varying the proportion of the total human-labeled data used to train the model. These subsets were again randomly sampled with data stratification. Predictions on *RB* are reported as $HL^{0\%}_{train}$, while $HL^{100\%}_{train}$ refers to $RB \rightarrow HL$. This ablation experiment was repeated 5 times to address training variation.

3. Results

3.1. Evaluation Metrics

Each model was evaluated for its ability to automatically extract findings from the reports, specifically by computing metrics on the negative and positive mentions of the findings. The performance was assessed by computing the F1 score for the positive and negative mention classes across 49 findings. These scores were then averaged to obtain the *macro* F1 score. The F1 score represents both precision and recall in one metric. It can be defined as

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN} \quad (1)$$

To address the class imbalance, the *weighted* F1 score is also reported, which weights the macro F1 score by the class support.

3.2. RegEx and Transfer Learning

When comparing the distribution of the F1 across 49 findings in Figure 5, most models performed better on the positive mention class. However, *BotXO* and *DanskBERT* showed similar distributions for both negative and positive findings. Table 2 shows that every machine learning model surpassed *RE* in terms of the macro F1 score, with the exception of *mBERT* on the negative mentions. The latter barely reached the performance of the *RE* method, but still showed an improvement in the other classes. Interestingly, despite being trained on the target domain data, *MeDa-BERT* did not perform as well as *BotXO* on both positive and negative mentions. The larger RoBERTa models *XLM* and *DanskBERT* performed slightly worse on positive mentions compared to BERT models; however, they achieved the biggest improvement over *RE* for negative mentions ($\Delta = +0.046$) and the

weighted F1 score for all findings ($\Delta = +0.077$). Looking at the standard deviation across every experiment, there was very little variance between the multiple training runs with different seeds, meaning that the selected training schedule was consistent.

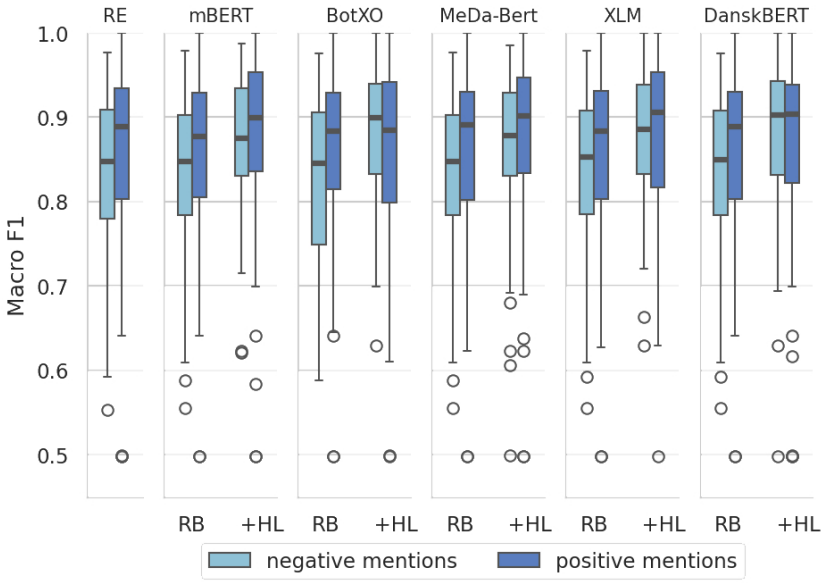

Figure 5. Distribution of macro F1 score across all 49 findings for the positive and negative mention classes for *RB* and $RB \rightarrow HL$.

Table 2. Macro F1 and weighted F1 scores for positive and negative mentions for the 49 findings in the *HL* test set, for $RB \rightarrow HL$ models. The scores are the average score and one standard deviation interval over 5 repeated $RB \rightarrow HL$ training runs: for each run, the same *RB* checkpoint was used for fine-tuning on *HL* labels, but with a different seed. Best score per metric is highlighted in bold.

	Positive F1	**Negative F1**	**Weighted F1**	
			All Findings	**Most Frequent**
RE	0.721	0.478	0.667	0.846
mBERT	0.742 ± 0.003	0.477 ± 0.008	0.732 ± 0.003	0.869 ± 0.001
BotXO	**0.745** ± 0.007	0.509 ± 0.012	0.737 ± 0.004	0.876 ± 0.002
MeDa-BERT	0.739 ± 0.005	0.480 ± 0.006	0.742 ± 0.003	0.873 ± 0.002
XLM	0.738 ± 0.007	0.498 ± 0.004	0.736 ± 0.004	**0.884** ± 0.001
DanskBERT	0.738 ± 0.011	**0.524** ± 0.008	**0.744** ± 0.007	0.882 ± 0.003

3.3. Most Frequent Findings Subanalysis

By restricting the number of tested findings to only the most frequent ones, there was a important improvement in the weighted F1 scores, as shown in Table 2. For example, the F1 score for frequent findings increased ($\Delta = +0.138$) for DanskBERT compared to that for all findings. This means that even if, on average, the model's average performance was lowered by the inclusion of underrepresented labels, its performance on the most frequent findings was still enhanced compared to *RE*. For the macro F1 scores of the positive and negative mentions for the most frequent findings, see Table A3 in Appendix C.

Figure 6 illustrates the performance of models on the individual most frequent findings. Most models tended to cluster around the same score when the class support was high enough, and the language models performed better than the *RE*. Exceptions arose with findings that were underrepresented, such as *Consolidation* in the positive mention class ($n = 54$) and *ChronicLungChanges* in the negative mention class ($n = 2$).

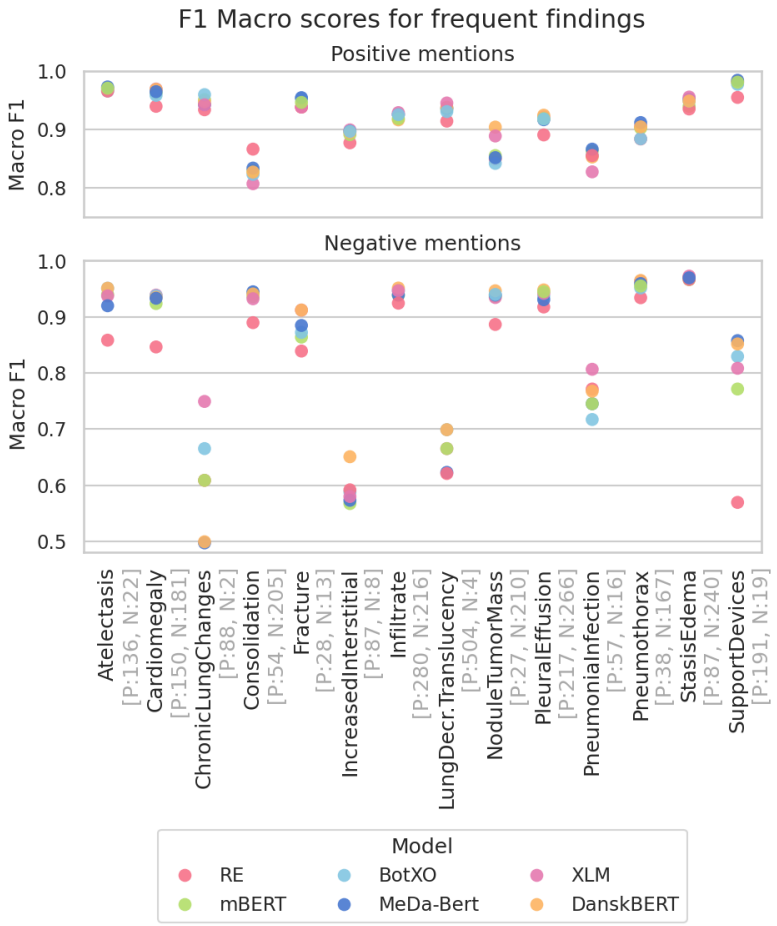

Figure 6. F1 scores of the most frequent findings. In square brackets, the label and class support in HL_{test}, where *P* stands for positive mentions and *N* stands for negative mentions.

3.4. Model Ensemble Subanalysis

Table 3 presents the results of the model ensemble experiments. In comparison to the single-model performance in Table 2, all of the ensembled models performed slightly worse on positive mention classification but improved on negative mention prediction. The only exception was *MeDa-BERT*, which consistently struggled with this class. Overall, the performance of the ensembled *DanskBERT* model improved the most compared to its unensembled counterpart for negative mention prediction ($\Delta = +0.193$). We noted that the ensembled *DanskBERT* model did not perform as well as XLM on positive mention classification ($\Delta = -0.016$), but it outperformed all models on negative mention classification.

Table 3. Macro F1 scores of the 49 findings for the *HL* test set for the $RB \rightarrow HL$ model ensembles, compared to *RE*. Best score per metric is highlighted in bold.

	Positive F1	**Negative F1**	**Weighted F1**
RE	0.721	0.478	0.667
mBERT	0.743	0.650	0.766
BotXO	0.726	0.693	0.763
MeDa-BERT	0.747	0.397	0.727
XLM	**0.756**	0.516	0.748
DanskBERT	0.740	**0.717**	**0.778**

Figure 7 shows the macro F1 scores of both the negative and positive mentions for each *k-fold*, compared against the fold average and the final result obtained through majority voting. The majority voting score for all models was higher than the average score across individual folds, indicating a substantial enhancement in performance through ensembling. The results for *XLM* and *MeDa-Bert* were clustered together, while the rest of the models showed more variation. Every model, except for *MeDa-Bert*, improved (became closer to the top-right corner) with majority voting compared to just the fold average, especially for negative mentions, demonstrating that model ensembles can improve the robustness of predictions.

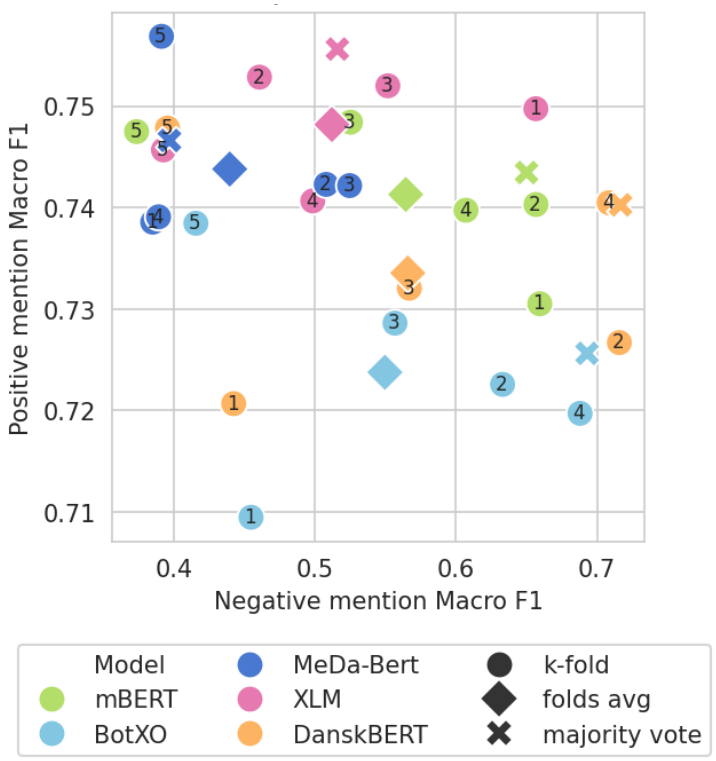

Figure 7. Positive and negative mention macro F1 scores for each *k-fold* trained for the model ensembles, including the averages across folds and the scores of the models' ensembles obtained through majority voting fold predictions.

3.5. Data Ablation Subanalysis

Figure 8 illustrates the impact of increasing the amount of human-labeled training data to better demonstrate how much expert-labeled data is necessary to achieve these improvements over the rule-based system. For all findings (Figure 8a), larger training sets slightly improved the accuracy of negative mentions, the most challenging class to learn. However, this improvement was accompanied by a slight decrease in performance for positive mentions. The $HL_{train}^{50\%}$ training set demonstrated a marked drop in performance across all five runs. We can find no explanation to support this unexplained sudden drop in performance, which persisted over every model and on different training runs. Only *BotXO* and *mBERT*'s performance gradually recovered, but with more variation. In the right part of the figure (Figure 8b), the evaluations were limited to the most frequent findings, which resulted in little to no variance among training runs, but they showed a slight decrease in performance on $HL_{train}^{50\%}$. More detailed results on data ablation are reported in Table A5 (Appendix C).

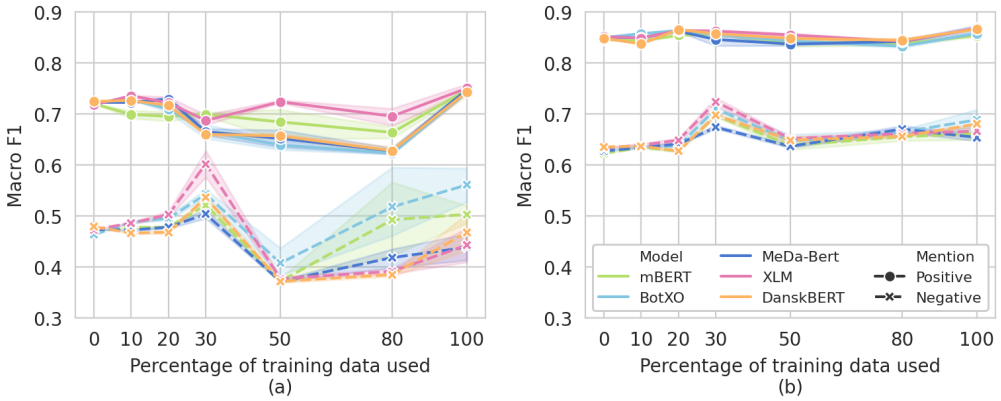

Figure 8. Data ablation study on $RB \rightarrow HL$, showing F1 score for positive (•) and negative (×) mentions across five model training runs. Scores on RB are reported as $HL^{0\%}_{train}$, while $HL^{100\%}_{train}$ refers to $RB \rightarrow HL$. (**a**) Results on all findings and (**b**) on most frequent findings.

Lastly, Table 4 summarizes the results presented, including those of experiments on models trained exclusively on RB and HL. If no rule-based annotations were available, the performance degraded drastically, to the point that RE performed better than machine learning models trained on only 1600 samples.

Table 4. Weighted F1 score of the 49 findings for the HL test set, summarizing the results obtained using different training sets and strategies. Extended results are available in Tables A1–A4 in Appendix C.

	RB	HL	$RB \rightarrow HL$	**Ensemble**
RE	0.667	-	-	-
mBERT	0.702	0.566	0.732	0.766
BotXO	0.700	0.511	0.737	0.763
MeDa-BERT	0.704	0.540	0.742	0.727
XLM	0.701	0.532	0.736	0.748
DanskBERT	0.707	0.535	0.744	**0.778**

4. Discussion

The aim of this study was to classify Danish chest X-ray text reports by the abnormalities mentioned by radiologists. An ensemble of predictions from the larger model trained on more Danish data achieved the best overall classification performance, particularly in negative mention detection. Machine learning models trained on a combination of automatically and manually extracted labels outperformed classic rule-based methods. Models pre-trained on Danish slightly outperformed multi-language models. Surprisingly, the model trained on more high-quality Danish medical data did not benefit from continued pre-training on the target domain data. Taken together, these results are encouraging for languages that lack good, open-weight pre-trained models, suggesting that multi-lingual models may be used instead to achieve a similar performance. However, it may still be necessary to adapt such a multi-lingual model to each language because it is not straightforward to train a single classification model for many languages [29].

For comparison to English methods, Smit et al. [17] reported a weighted F1 score of 0.798 for 14 findings on reports written in English and a radiologist benchmark of 0.805. While it cannot be directly compared, the proposed Danish method exceeded the expectation of matching previous work results, as it achieved an F1 score of 0.882 on a similar label set, although without an evaluation of the uncertainty extraction.

The labeling protocol plays a crucial role when evaluating performance against human annotators. Chen et al. claim that implementing a hierarchical taxonomy for label extraction improves the labeling accuracy and reduces missing annotations [30]. However, classification problems get harder when the set of classes increases [31]. All the tested machine learning models, when limited to the most frequent findings, showed an improvement in the macro F1 score, demonstrating that including less frequent abnormalities does not imply a drastic drop in performance on the more frequent findings. However, to ensure a better result for uncommon findings, more samples are required. When considering a large and diverse set of findings, model ensembles showed an increase in the robustness and generalization of the predictions. The expert-annotated data ablation study further emphasized the importance of the quality and distribution of training dataset labels over the dataset size. This result suggests that there might be less need for large sets of annotated samples than expected if previous training on low-quality data is performed. A curated distribution of findings may be more beneficial than increasing the dataset's size.

Some challenges presented by this task were addressed by implementing and testing a strategy for the creation of annotation tools for the classification of radiology reports; creating or collecting different NLP methods for Danish and medical-domain Danish; and addressing the variation in the performance of tools that were created using domain-specific data. The limitations of this study include the difficult access to medical datasets due to restricted access to radiology reports and the corresponding annotations. Since the focus of this study was on a single dataset, it potentially limits the generalization of the reported findings. Additionally, while this dataset did not necessitate the use of larger language model context windows for analysis, reports from different clinical settings may be longer and place critical information towards the end. No class-balancing techniques were deployed in this study as the dataset was large enough to reflect the final target distribution, but these may be advised when the distribution of classes and labels is highly irregular and not representative.

In summary, the most effective model was the 125 million-parameter *DanskBERT* model pre-trained on general-purpose text in our target language, with predictions determined through majority voting from an ensemble of five such models. Methods developed for English can also be applied to another language, such as Danish, but models benefit from a bigger target language pre-training dataset. The combination of good-quality RegEx rules and large language models proved essential to solve this task, and accessible, open-weight, multi-language models are an adequate solution in the absence of language-specific models.

5. Conclusions

This study compared regular expression rules and machine learning models for detecting medical findings within Danish radiology reports. The task was challenging as the available methods have been developed mainly for English, and no such tools are available for Danish. In general, large language models outperformed *RegEx* rules, particularly when models were pre-trained on more Danish text, especially in capturing negative mentions. Ensemble methods improved the model generalization. In this task, a detailed labeling protocol did not hurt the performance of machine learning models on the most frequent abnormalities. However, to achieve optimal performance on underrepresented findings, the dataset size and its diversity must be prioritized. While this study is limited to a single European language, future work might verify if the techniques developed for this study can be applied to other Germanic languages or other Indo-European languages. Exploring the integration of text-based findings with image classifiers could offer valuable insights into potential performance improvements for the development of AI radiology solutions.

Author Contributions: Conceptualization, A.S. and D.E.; methodology, A.S., R.B., M.F. and D.E.; software, A.S., R.B. and M.F.; validation, A.S.; formal analysis, A.S., R.B. and M.F.; investigation, A.S., R.B. and M.F.; resources, L.M.P., S.I., D.L. and M.B.N.; data curation, A.S., L.M.P., S.I., M.F. and M.B.N.; writing—original draft preparation, A.S., L.M.P. and D.E.; writing—review and editing, A.S., L.M.P., S.I., R.B., M.F., D.L., M.B.N. and D.E.; visualization, A.S. and L.M.P.; supervision, M.B.N. and D.E.; project administration, S.I. and D.E.; funding acquisition, M.B.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innovation Fund Denmark grant number 0176-00013B.

Institutional Review Board Statement: Ethical approval was obtained on 11 May 2022 from the Regional Council for Region Hovedstaden (R-22017450).

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available due to legal restrictions on the distribution medical records data.

Conflicts of Interest: Author Silvia Ingala was employed by the company Cerebriu A/S, Authors Rasmus Bonnevie and Marco Fraccaro were employed by the company Unumed Aps. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Data Description

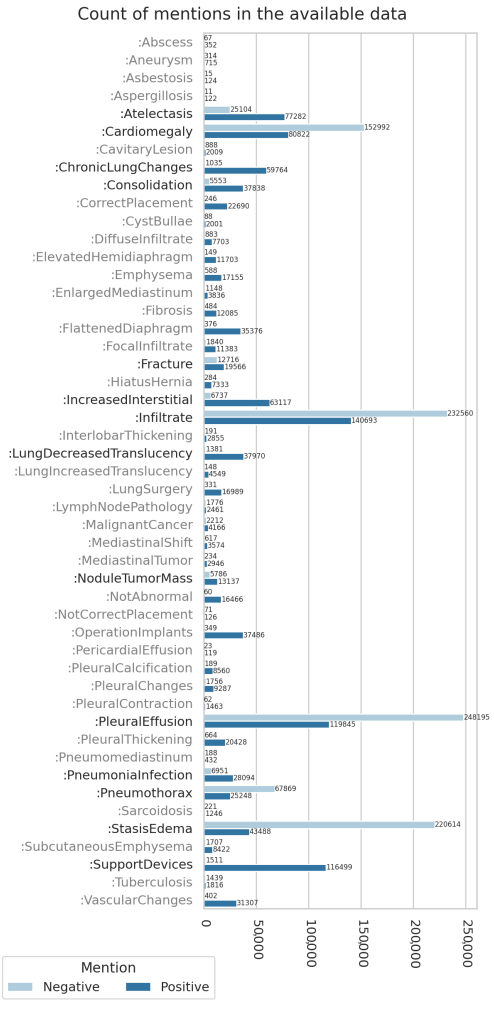

Figure A1. Distribution of labels in the dataset (550,233 samples in total), including annotations from *RB* or *HL* when available. In black, the most frequent findings.

Listing A1. Hierarchical tree structure defined for the annotation of chest X-ray reports, translated from Danish into English for readability.

```
Root
├── Abscess
├── Asbestosis
├── Aspergillosis
├── Cardiomegaly
├── ChronicLungChanges
├── ElevatedHemidiaphragm
├── EnlargedMediastinum
│   ├── LymphNodePathology
│   ├── MediastinalTumor
│   └── VascularChanges
│       └── Aneurysm
├── Fibrosis
├── FlattenedDiaphragm
├── Fracture
├── HiatusHernia
├── Infiltrate
│   ├── CavitaryLesion
│   ├── Consolidation
│   ├── DiffuseInfiltrate
│   ├── FocalInfiltrate
│   └── NoduleTumorMass
├── LungDecreasedTranslucency
│   ├── Atelectasis
│   ├── IncreasedInterstitial
│   ├── InterlobarThickening
│   └── PleuralEffusion
├── LungIncreasedTranslucency
│   ├── CystBullae
│   ├── Emphysema
│   └── Pneumothorax
├── LungSurgery
├── MalignantCancer
├── MediastinalShift
├── NotAbnormal
├── OperationImplants
├── PericardialEffusion
├── PleuralChanges
│   ├── PleuralContraction
│   └── PleuralThickening
│       └── PleuralCalcification
├── Pneumomediastinum
├── PneumoniaInfection
├── Sarcoidosis
├── StasisEdema
├── SubcutaneousEmphysema
├── SupportDevices
│   ├── CorrectPlacement
│   └── NotCorrectPlacement
└── Tuberculosis
```

Appendix B. Implementation Details

Appendix B.1. Regular Expressions

The RegEx rules for *RE* were designed based on the manual review of sentences from the training reports. These rules can be used as a formal shorthand for writing patterns that could capture textual variations at the level of spelling, white space, synonyms, and typos, while remaining legible to subject matter experts. As such, generators like $*$ ("$*$" matches zero or more occurrences of the preceding element and can match many variations infinitely) and $+$ ("$+$" matches one or more occurrences of the preceding element) were excluded, and all RegEx patterns are finite. The rules were designed by identifying phrases that were repeated in the reports or had the character of medical observations and then passed to expert radiologists for coding, with examples drawn from the text. To avoid having to manually account for conjugation, the patterns were expanded to full strings, and both pattern strings and report sentences were tokenized (dividing the text into individual words or subwords (tokens) for easier analysis) and lemmatized (reducing words to their base or root form to standardize variations, i.e., `[training,trains,trained]` $\rightarrow$ `train`) to look for matches. If patterns overlap, the longer match is chosen, which is typically the most specific. To account for negations, Danish negation patterns were designed to be used with the NegEx algorithm by Chapman et al. [24].

By design, the rule-based labeler only assigns a single label per match. In a hierarchical setting, such a labeling strategy is incomplete because multiple related labels could be inferred from any one observation. Concretely, the hierarchy is interpreted as an ontology of nested concepts: a mention of *DiffuseInfiltrate* is also a member of the parent class *Infiltrate* and acquires a positive class label. Another example is the sentence "`The patient has diffuse infiltrates in the left lung.`", which is a direct *DiffuseInfiltrate* finding or an indirect *Infiltrate* finding of the positive class. Given direct memberships only, the indirect labels can be expanded by propagating any positive label up to all of its ancestors. A negative mention cannot be propagated up, as the patient might have a positive finding of a more generic type, but it can be propagated down instead, i.e., if the patient does not have *Infiltrate*, they also do not have *DiffuseInfiltrate*. If there is a conflict, the positive label wins out.

Appendix B.2. Large Language Models

Instead of models trained only on English, the models used to solve this task were selected based on their performance on the ScandEval Benchmark [32], which tests the capabilities of various large language models for solving tasks in Scandinavian languages. The selected models were the top-performing, which were fine-tuned on multi-language datasets (including Danish) or specifically on Danish. All the models are open-weight and available through the HuggingFace Transformers library. Thus, different BERT$_{BASE}$ ($BASE$ commonly refers to the smallest available size of a given family of LLMs) and RoBERTa$_{BASE}$ [33] architectures were picked, given their general suitability for this task [34].

mBERT, short for BERT base multi-lingual uncased [18], is a model with 110 million parameters that was trained on a large corpus of 104 languages, including Danish, for the tasks of masked language modeling and next sentence prediction. *BotXO*, short for the Danish BERT model by BotXO [35], is a BERT model trained from scratch on 9.5 GB of Danish text. This model was expected to achieve a higher performance compared to *mBERT*, as *BotXO* was trained in the target language. *MeDa-BERT*, short for Danish medical BERT [36], was initialized from *BotXO*, and it was further fine-tuned on a Danish medical corpus of 123 M tokens. Given that medical text was the language domain of the task, an even better performance was expected. *XLM*, short for XLM-RoBERTa base [37], is the multi-lingual version of RoBERTa, which outperforms multi-lingual BERT (*mBERT*) on a variety of cross-lingual benchmarks. This model is larger than BERT, having 125 million parameters. It was trained on 2.5 TB of data in 100 languages, including Danish text. *DanskBERT*, developed by Snæbjarnarson et al. [38], is a RoBERTa base model further fine-tuned on 2.2 GiB of Danish text. Like with *BotXO*, a better performance than that of the multi-lingual base model *XLM* was expected.

Appendix C. Detailed Results

The following results refer to a single training experiment. In brackets is the variance defined as the standard deviation from the mean for the set of 49 findings.

Table A1. Macro F1 score for the 49 findings on the *HL* test set for models trained only on the *HL* training set. In bold, the absolute best score and the best score for the transformer models are underlined. Compared to the *RE* method, no transformer alone was able to match its string-matching performance.

	Positive F1	**Negative F1**	**Weighted F1**
RE	**0.721** (0.65–0.79)	**0.478** (0.40–0.55)	**0.667** (0.61–0.73)
mBERT	0.491 (0.39–0.59)	0.308 (0.20–0.41)	0.566 (0.48–0.65)
BotXO	0.455 (0.35–0.56)	0.279 (0.18–0.38)	0.511 (0.42–0.60)
MeDa-BERT	0.447 (0.34–0.55)	0.326 (0.22–0.43)	0.540 (0.45–0.63)
XLM	0.471 (0.37–0.57)	0.299 (0.19–0.40)	0.532 (0.44–0.62)
DanskBERT	0.466 (0.36–0.57)	0.309 (0.21–0.41)	0.535 (0.44–0.63)

Table A2. Macro F1 score for the 49 findings on the *HL* test set for models trained only on the *RB* training set. Best score per metric is highlighted in bold.

	Positive F1	**Negative F1**	**Weighted F1**
RE	0.721 (0.65–0.79)	0.478 (0.40–0.55)	0.667 (0.61–0.73)
mBERT	0.720 (0.65–0.79)	0.471 (0.40–0.55)	0.702 (0.64–0.76)
BotXO	0.722 (0.65–0.79)	0.463 (0.39–0.54)	0.700 (0.64–0.76)
MeDa-BERT	0.723 (0.65–0.79)	0.474 (0.40–0.55)	0.704 (0.64–0.77)
XLM	0.718 (0.65–0.79)	0.473 (0.40–0.55)	0.701 (0.64–0.76)
DanskBERT	**0.725** (0.65–0.80)	**0.479** (0.40–0.55)	**0.707** (0.64–0.77)

Table A3. Macro F1 score for the 49 findings on the *HL* test set for $RB \rightarrow HL$ models, but reporting only the averages over the 14 most frequent findings. Best score per metric is highlighted in bold.

	Positive F1	**Negative F1**	**Weighted F1**
RE	0.848 (0.80–0.89)	0.635 (0.46–0.81)	0.844 (0.80–0.89)
mBERT	0.857 (0.80–0.91)	0.682 (0.50–0.87)	0.874 (0.83–0.92)
BotXO	0.860 (0.80–0.92)	**0.721** (0.57–0.87)	0.881 (0.84–0.93)
MeDa-BERT	**0.877** (0.83–0.92)	0.663 (0.46–0.87)	**0.883** (0.84–0.93)
XLM	0.868 (0.81–0.92)	0.653 (0.42–0.88)	0.879 (0.82–0.94)
DanskBERT	0.867 (0.82–0.92)	0.679 (0.47–0.88)	**0.883** (0.84–0.93)

Table A4. Macro F1 score for the 49 findings on the *HL* test set for $RB \rightarrow HL$ model ensembles, compared to *RE*. Best score per metric is highlighted in bold.

	Positive F1	**Negative F1**	**Weighted F1**
RE	0.721 (0.65–0.79)	0.478 (0.40–0.55)	0.667 (0.61–0.73)
mBERT	0.743 (0.67–0.81)	0.650 (0.57–0.73)	0.766 (0.71–0.82)
BotXO	0.726 (0.65–0.80)	0.693 (0.62–0.77)	0.763 (0.71–0.82)
MeDa-BERT	0.747 (0.67–0.82)	0.397 (0.28–0.51)	0.727 (0.66–0.79)
XLM	**0.756** (0.68–0.83)	0.516 (0.41–0.62)	0.748 (0.69–0.81)
DanskBERT	0.740 (0.67–0.81)	**0.717** (0.63–0.80)	**0.778** (0.72–0.83)

Table A5. Results of the $RB \rightarrow HL$ models trained on only a percentage of HL_{train}, averaged over 5 training runs.

		10%	20%	30%	50%	80%
mBERT	Negative F1	0.478	0.478	0.521	0.372	0.493
	Positive F1	0.699	0.695	0.699	0.684	0.664
	Weighted F1	0.690	0.692	0.710	0.685	0.696
BotXO	Negative F1	0.488	0.495	0.543	0.408	0.518
	Positive F1	0.728	0.709	0.666	0.639	0.623
	Weighted F1	0.711	0.708	0.695	0.666	0.677
MeDa-BERT	Negative F1	0.472	0.478	0.504	0.372	0.418
	Positive F1	0.723	0.729	0.665	0.652	0.628
	Weighted F1	0.712	0.715	0.702	0.676	0.673
XLM	Negative F1	0.486	0.503	0.602	0.377	0.391
	Positive F1	0.736	0.722	0.687	0.724	0.695
	Weighted F1	0.722	0.720	0.722	0.711	0.700
DanskBERT	Negative F1	0.466	0.468	0.536	0.372	0.385
	Positive F1	0.726	0.716	0.660	0.657	0.628
	Weighted F1	0.703	0.705	0.699	0.675	0.665

Appendix D. Negative Results

All machine learning methods outperformed *RE* when trained on the *RB* data. We conducted an initial experiment using a BERT model fine-tuned on *RB* as the base model instead, so that the model was then tuned on the labels predicted by the BERT model instead of *RB*. The model failed to converge, thus leading to very poor results. This is reported in Table A6 as a negative result. A manual inspection of the predictions on the human-labeled set revealed that DanskBERT(ML_{RB}) overestimated the number of mentioned findings, both positive and negative. Finally, DanskBERT($ML_{RB} \rightarrow HL$) broke completely by always predicting the negative mention class for every finding except "NotAbnormal", which was always classified as a "positive mention".

Table A6. Macro F1 score for the DanskBERT model trained on the labels produced by DanskBERT trained on *RB*, named DanskBERT(ML_{RB}), and how it compares to that of $RB \rightarrow HL$ and the same model further fine-tuned on *HL* (DanskBERT($ML_{RB} \rightarrow HL$)).

	Positive F1	Negative F1	Weighted F1
DanskBERT($RB \rightarrow HL$)	0.72	0.65	0.76
DanskBERT(ML_{RB})	0.04	0.30	0.28
DanskBERT($ML_{RB} \rightarrow HL$)	0.003	0.94	0.87

References

1. Ait Nasser, A.; Akhloufi, M.A. A Review of Recent Advances in Deep Learning Models for Chest Disease Detection Using Radiography. *Diagnostics* **2023**, *13*, 159. [CrossRef] [PubMed]
2. Rajpurkar, P.; Irvin, J.; Ball, R.L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P.; et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **2018**, *15*, e1002686. [CrossRef] [PubMed]
3. Esteva, A.; Chou, K.; Yeung, S.; Naik, N.; Madani, A.; Mottaghi, A.; Liu, Y.; Topol, E.; Dean, J.; Socher, R. Deep learning-enabled medical computer vision. *NPJ Digit. Med.* **2021**, *4*, 5. [CrossRef] [PubMed]
4. Rädsch, T.; Reinke, A.; Weru, V.; Tizabi, M.D.; Schreck, N.; Kavur, A.E.; Pekdemir, B.; Roß, T.; Kopp-Schneider, A.; Maier-Hein, L. Labelling instructions matter in biomedical image analysis. *Nat. Mach. Intell.* **2023**, *5*, 273–283. [CrossRef]
5. Wang, H.; Jin, Q.; Li, S.; Liu, S.; Wang, M.; Song, Z. A comprehensive survey on deep active learning in medical image analysis. *Med. Image Anal.* **2024**, *95*, 103201. [CrossRef] [PubMed]
6. Chng, S.Y.; Tern, P.J.W.; Kan, M.R.X.; Cheng, L.T.E. Automated labelling of radiology reports using natural language processing: Comparison of traditional and newer methods. *Health Care Sci.* **2023**, *2*, 120–128. [CrossRef]

7. Pereira, S.C.; Mendonça, A.M.; Campilho, A.; Sousa, P.; Teixeira Lopes, C. Automated image label extraction from radiology reports—A review. *Artif. Intell. Med.* **2024**, *149*, 102814. [CrossRef] [PubMed]
8. Zingmond, D.; Lenert, L.A. Monitoring free-text data using medical language processing. *Comput. Biomed. Res.* **1993**, *26*, 467–481. [CrossRef] [PubMed]
9. George Hripcsak, M.; Carol Friedman, P.; Philip, O.; Alderson, M.; William DuMouchel, P.; Stephen, B.; Johnson, P.; Paul, D.; Clayton, P. Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing. *Ann. Intern. Med.* **1995**, *122*, 681–688. [CrossRef] [PubMed]
10. Jain, N.L.; Knirsch, C.A.; Friedman, C.; Hripcsak, G. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. In Proceedings of the AMIA Annual Fall Symposium, Washington, DC, USA, 26–30 October 1996; p. 542.
11. Peng, Y.; Wang, X.; Lu, L.; Bagheri, M.; Summers, R.; Lu, Z. NegBio: A high-performance tool for negation and uncertainty detection in radiology reports. *arXiv* **2017**, arXiv:cs.CL/1712.05898.
12. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv* **2019**. [CrossRef]
13. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3462–3471. [CrossRef]
14. Reis, E.P.; de Paiva, J.P.Q.; da Silva, M.C.B.; Ribeiro, G.A.S.; Paiva, V.F.; Bulgarelli, L.; Lee, H.M.H.; Santos, P.V.; Brito, V.M.; Amaral, L.T.W.; et al. BRAX, Brazilian labeled chest x-ray dataset. *Sci. Data* **2022**, *9*, 487. [CrossRef]
15. Nguyen, T.; Vo, T.M.; Nguyen, T.V.; Pham, H.H.; Nguyen, H.Q. Learning to diagnose common thorax diseases on chest radiographs from radiology reports in Vietnamese. *PLoS ONE* **2022**, *17*, e0276545. [CrossRef] [PubMed]
16. Wollek, A.; Hyska, S.; Sedlmeyr, T.; Haitzer, P.; Rueckel, J.; Sabel, B.O.; Ingrisch, M.; Lasser, T. German CheXpert Chest X-ray Radiology Report Labeler. *arXiv* **2023**, arXiv:cs.CL/2306.02777. [CrossRef]
17. Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A.Y.; Lungren, M.P. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. *arXiv* **2020**. [CrossRef]
18. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
19. Rehm, G.; Berger, M.; Elsholz, E.; Hegele, S.; Kintzel, F.; Marheinecke, K.; Piperidis, S.; Deligiannis, M.; Galanis, D.; Gkirtzou, K.; et al. European Language Grid: An Overview. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 3366–3380.
20. Li, D.; Pehrson, L.M.; Lauridsen, C.A.; Tøttrup, L.; Fraccaro, M.; Elliott, D.; Zając, H.D.; Darkner, S.; Carlsen, J.F.; Nielsen, M.B. The Added Effect of Artificial Intelligence on Physicians' Performance in Detecting Thoracic Pathologies on CT and Chest X-ray: A Systematic Review. *Diagnostics* **2021**, *11*, 2206. [CrossRef]
21. Li, D.; Pehrson, L.M.; Tøttrup, L.; Fraccaro, M.; Bonnevie, R.; Thrane, J.; Sørensen, P.J.; Rykkje, A.; Andersen, T.T.; Steglich-Arnholm, H.; et al. Inter- and Intra-Observer Agreement When Using a Diagnostic Labeling Scheme for Annotating Findings on Chest X-rays—An Early Step in the Development of a Deep Learning-Based Decision Support System. *Diagnostics* **2022**, *12*, 3112. [CrossRef]
22. Li, D.; Pehrson, L.M.; Bonnevie, R.; Fraccaro, M.; Thrane, J.; Tøttrup, L.; Lauridsen, C.A.; Butt Balaganeshan, S.; Jankovic, J.; Andersen, T.T.; et al. Performance and Agreement When Annotating Chest X-ray Text Reports—A Preliminary Step in the Development of a Deep Learning-Based Prioritization and Detection System. *Diagnostics* **2023**, *13*, 1070. [CrossRef]
23. Bustos, A.; Pertusa, A.; Salinas, J.M.; De La Iglesia-Vaya, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **2020**, *66*, 101797. [CrossRef] [PubMed]
24. Chapman, W.W.; Bridewell, W.; Hanbury, P.; Cooper, G.F.; Buchanan, B.G. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *J. Biomed. Inform.* **2001**, *34*, 301–310. [CrossRef] [PubMed]
25. von der Mosel, J.; Trautsch, A.; Herbold, S. On the Validity of Pre-Trained Transformers for Natural Language Processing in the Software Engineering Domain. *IEEE Trans. Softw. Eng.* **2023**, *49*, 1487–1507. [CrossRef]
26. Sechidis, K.; Tsoumakas, G.; Vlahavas, I. On the stratification of multi-label data. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, 5–9 September 2011; Proceedings, Part III 22; Springer: Berlin/Heidelberg, Germany, 2011; pp. 145–158.
27. Dietterich, T.G. Ensemble methods in machine learning. In Proceedings of the International Workshop on Multiple Classifier Systems, Cagliari, Italy, 21–23 June 2000; pp. 1–15.
28. Mohammed, A.; Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud-Univ.-Comput. Inf. Sci.* **2023**, *35*, 757–774. [CrossRef]

29. Kær Jørgensen, R.; Hartmann, M.; Dai, X.; Elliott, D. mDAPT: Multilingual Domain Adaptive Pretraining in a Single Model. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November 2021; pp. 3404–3418. [CrossRef]
30. Chen, H.; Miao, S.; Xu, D.; Hager, G.D.; Harrison, A.P. Deep Hierarchical Multi-label Classification of Chest X-ray Images. In Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning, PMLR, London, UK, 8–10 July 2019; Volume 102, pp. 109–120.
31. Del Moral, P.; Nowaczyk, S.; Pashami, S. Why is multiclass classification hard? *IEEE Access* **2022**, *10*, 80448–80462. [CrossRef]
32. Nielsen, D. ScandEval: A Benchmark for Scandinavian Natural Language Processing. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), Tórshavn, Faroe Islands, 22–24 May 2023; pp. 185–201.
33. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
34. Dai, X.; Chalkidis, I.; Darkner, S.; Elliott, D. Revisiting Transformer-based Models for Long Document Classification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 7212–7230. [CrossRef]
35. BotXO, CertainlyIO. Danish BERT. 2020. Available online: https://github.com/botxo/nordic_bert (accessed on 4 April 2024).
36. Pedersen, J.; Laursen, M.; Vinholt, P.; Savarimuthu, T.R. MeDa-BERT: A medical Danish pretrained transformer model. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), Tórshavn, Faroe Islands, 22–24 May 2023; pp. 301–307.
37. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv* **2019**, arXiv:abs/1911.02116.
38. Snæbjarnarson, V.; Simonsen, A.; Glavaš, G.; Vulić, I. Transfer to a Low-Resource Language via Close Relatives: The Case Study on Faroese. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), Tórshavn, Faroe Islands, 22–24 May 2023.