JMIR Publications
The Leading eHealth Publisher

Enhancing BERT with Frame Semantics to Extract Clinically Relevant Information from German Mammography Reports: Algorithm Development and Validation

Daniel Reichenpfader, Jonas Knupp, Sandro von Däniken, Roberto Gaio, Fabio Dennstädt, Grazia Maria Cereghetti, André Sander, Hans Hiltbrunner, Knud Nairz, Kerstin Denecke

Submitted to: Journal of Medical Internet Research
on: November 07, 2024

Preprint JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint JMIR Publications

Enhancing BERT with Frame Semantics to Extract Clinically Relevant Information from German Mammography Reports: Algorithm Development and Validation

Daniel Reichenpfader[1, 2]; Jonas Knupp[3]; Sandro von Däniken[4]; Roberto Gaio[5]; Fabio Dennstädt[5]; Grazia Maria Cereghetti[4]; André Sander[3]; Hans Hiltbrunner[4]; Knud Nairz[4]; Kerstin Denecke[1]

[1]Institute for Patient-Centered Digital Health School of Engineering and Computer Science Bern University of Applied Sciences Biel/Bienne CH
[2]PhD School of Life Sciences Faculty of Medicine University of Geneva Geneva CH
[3]ID Suisse AG St. Gallen CH
[4]Department of Diagnostic, Interventional, and Pediatric Radiology Inselspital Bern University Hospital and University of Bern Bern CH
[5]Department of Radiation Oncology Inselspital Bern University Hospital and University of Bern Bern CH

Corresponding Author:
Daniel Reichenpfader
Institute for Patient-Centered Digital Health
School of Engineering and Computer Science
Bern University of Applied Sciences
Quellgasse 21
Biel/Bienne
CH

preprint JMIR Publications

Abstract

Background: Structured reporting is essential for improving the clarity and accuracy of radiological information. Despite its benefits, the European Society of Radiology (ESR) notes that it is not widely adopted. This raises the need for automatic methods to extract relevant information from unstructured radiology reports and thereby create structured reports automatically.

Objective: This study explores to combine a Bidirectional Encoder Representations from Transformers (BERT) architecture with the linguistic concept of frame semantics to extract and normalize information from free-text mammography reports.

Methods: After creating an annotated corpus of 210 German reports for fine-tuning, we generate several BERT-model variants by applying three pre-training strategies on hospital data. Afterwards, a fact extraction pipeline is built, comprising an extractive question answering model and a sequence labelling model. We evaluate all model variants quantitatively using common evaluation metrics (model perplexity, squad_v2, seqeval) and perform qualitative evaluation of the whole pipeline by clinicians on a manually created synthetic dataset of 21 reports.

Results: Our system is capable of extracting 14 fact types and 40 entities from the clinical findings section of mammography reports. Further pre-training on hospital data reduced model perplexity, although not having significant impact on the two downstream-tasks. We achieved averaged F1 scores of >90 % and >80 % for question answering and sequence labelling, respectively. Qualitative evaluation of the pipeline based on synthetic data shows overall precision of 96.1 % and 99.6 % for facts and entities, respectively.

Conclusions: The proposed BERT-based framework incorporating frame semantics effectively extracts structured information from unstructured radiology reports. This system shows promise for advancing automated structured reporting in radiology, supporting improved clarity and usability of radiological data.

(JMIR Preprints 07/11/2024:68427)
DOI: https://doi.org/10.2196/preprints.68427

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint
JMIR Publications

Original Manuscript

Preprint JMIR Publications

Original Paper

Daniel Reichenpfader[1], Jonas Knupp[2], Sandro von Däniken[3], Roberto Gaio[4], Fabio Dennstädt[4], Grazia Maria Cereghetti[3], André Sander[2], Hans Hiltbrunner[3], Knud Nairz[3], and Kerstin Denecke[1]

[1] Institute for Patient-Centered Digital Health, Bern University of Applied Sciences, Biel/Bienne, Switzerland

[2] ID Suisse AG, St. Gallen, Switzerland

[3] Department of Diagnostic, Interventional, and Pediatric Radiology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

[4] Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

Corresponding author: Daniel Reichenpfader, Quellgasse 21, 2502 Biel/Bienne, Switzerland. Phone: +41 31 848 60 93. Email: daniel.reichenpfader@bfh.ch.

Enhancing BERT with Frame Semantics to Extract Clinically Relevant Information from German Mammography Reports: Algorithm Development and Validation

Abstract

Background: Structured reporting is essential for improving the clarity and accuracy of radiological information. Despite its benefits, the European Society of Radiology (ESR) notes that it is not widely adopted. This raises the need for automatic methods to extract relevant information from unstructured radiology reports and thereby create structured reports automatically.

Objective: This study explores to combine a Bidirectional Encoder Representations from Transformers (BERT) architecture with the linguistic concept of frame semantics to extract and normalize information from free-text mammography reports.

Methods: After creating an annotated corpus of 210 German reports for fine-tuning, we generate several BERT-model variants by applying three pre-training strategies on hospital data. Afterwards, a fact extraction pipeline is built, comprising an extractive question answering model and a sequence labelling model. We evaluate all model variants quantitatively using common evaluation metrics (model perplexity, squad_v2, seqeval) and perform qualitative evaluation of the whole pipeline by clinicians on a manually created synthetic dataset of 21 reports.

Results: Our system is capable of extracting 14 fact types and 40 entities from the clinical findings section of mammography reports. Further pre-training on hospital data reduced model perplexity, although not having significant impact on the two downstream-tasks. We achieved averaged F1 scores of >90 % and >80 % for question answering and sequence labelling, respectively. Qualitative evaluation of the pipeline based on synthetic data shows overall precision of 96.1 % and 99.6 % for facts and entities, respectively.

Conclusions: The proposed BERT-based framework incorporating frame semantics effectively extracts structured information from unstructured radiology reports. This system shows promise for advancing automated structured reporting in radiology, supporting improved clarity and usability of radiological data.

Keywords: radiology; information extraction; mammography; large language models; structured reporting; template filling; annotation; quality control; natural language processing

Preprint JMIR Publications

Introduction

Background

Radiology reports serve as a critical connector between medical imaging and patient care by condensing radiologists' findings and interpretive insights into written text to inform physicians in charge. Historically, these reports have typically been narrative medical letters. While synoptic/structured reporting has been shown to have several advantages and is therefore of increasing interest, most radiological reports are still presented in a narrative format [1]. This narrative style, while commonly rich in detail, often lacks the standardization seen in other areas of medical documentation. E.g., laboratory reports are renowned for their ease of creation, enhanced standardization, and immediate comprehensibility [2,3].

Ideally, radiology reports would combine the depth and flexibility of narrative information with the clarity and structure of lab reports, allowing for a quickly comprehensible, easy, and unambiguous use for referring physicians. However, the style of radiology reports reflects the conflicting priorities of coping with high-throughput and standardized processes in radiology departments and providing individualized and patient-centered diagnostic information [4].

Structured reporting in radiology is defined as the systematic approach to creating reports using standardized language and formats, which can include templates, checklists, and hyperlinks [5]. However, the creation of templates for structured reporting is a meticulous process, and the supporting software platforms are struggling with accommodating the complexity and nuance of individual cases [6]. For referrers, the advantages of structured reporting are manifold, since the presentation of information in a structured format enhances clarity, reduces ambiguity, and thereby facilitates decision-making [4]. For radiologists, structured reporting can improve report quality, consistency, clarity, and completeness, potentially leading to a reduction in reporting errors and ultimately in higher customer satisfaction [7].

One of the most prominent guidelines for standardizing reporting in radiology is the Breast Imaging Reporting and Data System (BI-RADS), developed by the American College of Radiology (ACR) [8]. BI-RADS provides a framework for describing mammographic findings, categorization of results, and recommendation of follow-up actions thereby aiding in decisions about management of breast cancer patients [9]. The BI-RADS lexicon includes standardized descriptors for breast lesions, assessment categories that stratify the risk of malignancy, and management recommendations, all of which contribute to a more consistent approach to breast imaging interpretation and reporting [9]. Although BI-RADS provides a ruleset for reporting, mammography reports generally consist of semi-structured sections of free-text. Automatically extracting relevant information from these free-text sections to produce a structured report has the potential to improve reporting quality and eventually treatment outcome. An opportunity to facilitate this process would be the application of large language models (LLMs), a deep-learning-based model architecture trained on extensive amount of text-data, introduced in 2017 by Vaswani et al. [10].

LLMs can be separated into encoder-based, sequence-to-sequence and decoder-based models based on their architecture. While encoder-based models, e.g. following the BERT architecture [11], excel in tasks requiring understanding and extracting information, decoder-based (or generative) models including the GPT model family [12] , are designed to produce fluent, coherent text outputs. Sequence-to-sequence models, e.g. the T5 architecture, are best suited for tasks that involve generating new sentences based on a specific input [13]. LLMs continue to show impressive capabilities for realizing tasks related to natural language processing (NLP) in the medical domain

Preprint
JMIR Publications

[14,15]. They partly outperform clinicians in, e.g., medical summary generation [16], answering of clinical questions [17] and also extraction of structured information from clinical text [18,19]. However, commercial state-of-the-art models continue getting larger, requiring an increasing extent of scarce hardware resources and training data. Moreover, although large generative models currently show the best performance across many tasks and benchmarks, they are less explainable than, e.g., encoder-based architectures [20]. Further limitations include models ‘hallucinating’ wrong or misleading outputs [21] and restricted use of commercial models due to the sensitive nature of patient data [22]. Considering these limitations, the application of encoder-based LLMs for information extraction (IE) is of high interest for research in medical NLP due to their relatively small size and inherent output transparency. The possibility of reusing existing models and adapting them to the peculiarities of an institution – called further pre-training – additionally increases the potential of applying LLMs in radiology.

However, existing research on the specific task of information extraction from radiology reports based on LLMs is limited. Although during the last five years, several studies have been published, reporting quality and comparability of studies is impaired as shown by our previous work, a scoping review on studies describing information extraction from radiology reports [23]: It was shown that until August 2023, only pre-transformer and encoder-based models were applied. Moreover, LLMs might improve generalizability of IE approaches. Common open challenges included missing validation on external data and augmentation of methodologies.
The authors of a recent systematic review on specifically BERT-based NLP applications in radiology concluded that the BERT architecture shows the “potential to elevate diagnostic accuracy, accelerate generation of reports and optimize patient care” [24].

Regarding the specific task of structuring mammography reports, Saha et al. conducted a scoping review on NLP applied to breast cancer reports [25]: The authors show that NLP applications facilitate quality control in routine mammography and that LLMs offer distinct advantages due to their transfer learning capabilities and better performance. Prior to LLMs, rule-based approaches were most frequently used, as a systematic review on NLP-based extraction of cancer concepts from clinical notes showed [26].

Based on the existing literature, it becomes apparent that the number of extracted entities is usually limited and that the entity structure and types are proprietary for each IE project. To counteract these limitations, Steinkamp et al. introduced a method that enables clinicians themselves to define atomic, reusable and standardized types of clinically relevant information called “facts” in 2019 [27]. A fact always corresponds to a continuous text span, comprising pre-defined entities, called “anchors” (unique key word or phrase of a fact) and “modifiers” (optional, containing additional information). The authors based this information model on the linguistic concept of frame semantics, coined by Fillmore [28]. For implementation, they applied pre-transformer, deep-learning based methods (bidirectional gated recurrent units).

Frame semantics interprets words and phrases based on underlying conceptual structures, or "frames", which provide essential context. Each frame represents a scenario with specific roles and relationships, showing how certain words trigger a network of related meanings [28]. Applied to information extraction in radiology, frame semantics enables the organized mapping of clinically relevant data by linking key entities. This structure helps to align extracted information with real-world clinical situations, making it possible to standardize complex medical findings and adapt them across various reporting formats and domains.

Preprint JMIR Publications

Study goals and contribution

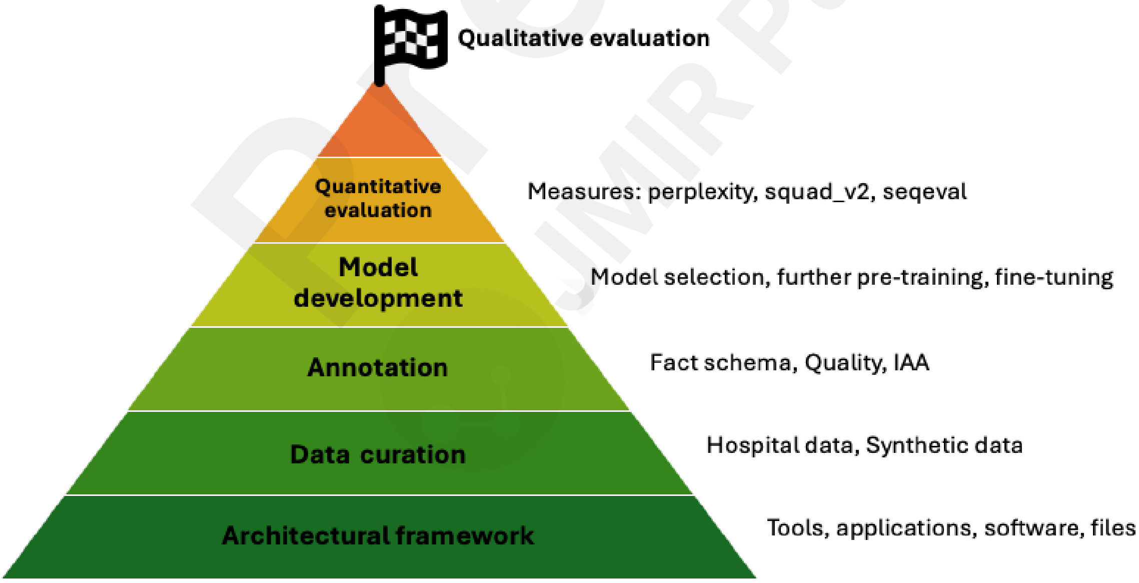
With this study, we want to investigate the application of encoder-based LLMs to build a pipeline that automatically extracts and normalizes clinically relevant information from mammography reports. Therefore, we conduct a novel approach combining the linguistic concept of frame semantics with recent LLMs. As a result, clinicians are enabled to define themselves what is relevant in a structured and reusable way. By further pre-training our models, we investigate whether model performance can be further improved.

Methods

Overview

This study implemented a previously introduced architectural framework for IE from clinical text [29]. In order to develop and evaluate our algorithms, our methodology further comprised the generation of real and synthetic datasets. Next, we designed and conducted an annotation sub-project to obtain high-quality annotations of the documents according to a set of facts and associated modifiers. This annotated dataset then allowed us to develop different variants of models to automatically extract the specified information. Finally, we evaluated the models quantitatively - using standardized evaluation metrics - as well as qualitatively. Qualitative evaluation was performed by two radiologists using the synthetic data. We made the source code as well as the trained models and containerized pipeline image available (see data availability statement). Model pre-training and fine-tuning was performed on a single Apple M1 Max chip. We describe each step of our methodology in more detail below and provide an illustrated overview of the mentioned steps in Figure 1.

Figure 1. Methodology overview. IAA: Inter-annotator-agreement


Qualitative evaluation
Quantitative evaluation
Measures: perplexity, squad_v2, seqeval
Model development
Model selection, further pre-training, fine-tuning
Annotation
Fact schema, Quality, IAA
Data curation
Hospital data, Synthetic data
Architectural framework
Tools, applications, software, files

System architecture

In Table 1, we provide an overview of the applied technologies to implement the RadEx framework [29]. The framework describes the necessary components and artifacts and includes various specifications. These specifications include the underlying information model, conducting the annotation process involving clinicians, for developing models to implement the IE tasks as well as

Preprint JMIR Publications

to use the models in production.

Table 1. Implementation of selected elements of the RadEx framework for information extraction from clinical texts [29]. UI: User Interface.

Framework component	Implementation
Fact schema definition UI	Atlassian Confluence [30]
Integration server	Custom Java application
Template filling	ID MACS® terminology server [31]
Inference Server	BentoML [32]
Library	Python/ Java, Github Packages [33]
Annotation generator	Custom Java application
Annotation Tool	INCEpTION [34]
Report preprocessor	Custom Java application
Annotation analyzer	INCEpTION, Custom Java application
Type system converter	Custom Java application
Model repositories	HuggingFace transformers library [35]

Dataset

We created separate datasets for self-supervised pre-training of LLM variants ("pre-training corpus") and for subsequently fine-tuning these variants for the task of IE("fine-tuning corpus"), see chapter "Model development". Both datasets were pre-processed before application.

For the pre-training corpus, anonymized, raw hospital data provided as multiple .xlsx files were merged into one .csv file and encoding errors (e.g., german umlauts) were fixed using a custom python script. Next, the file was symmetrically encrypted to allow pre-training on high performance computing infrastructure outside the hospital.

The creation of the fine-tuning corpus of mammography reports also comprised merging of raw data, stratification as well as transforming data into the required format for annotation: The first step, data merging and cleaning, comprised merging raw data provided as multiple .xlsx files into one .csv file, fixing of encoding errors, removal of duplicates and non-mammography reports as well as tokenization and sentence splitting. Tokenization and sentence splitting was based on the open-source library spaCy [36], augmented by a proprietary list of abbreviations and their associated expansions. Next, the reports were stratified according to the BI-RADS level of each report and split into three datasets for annotation guideline development (DS.A), modelling (DS.B) and validation (DS.C). Each dataset contained an equal number of reports from each BI-RADS category, sampled by maximum variance within each category. Last, each report was converted into the specified format (Common analysis structure, CAS) according to the architectural framework.

For qualitative evaluation, 21 synthetic reports were created by a radiology resident and verified by a senior radiologist. This synthetic dataset comprised three reports per BI-RADS category (0-6). A sample of such a report is shown in Textbox 1. We used this dataset for qualitative evaluation and made it publicly available (see data availability statement).

Textbox 1. Example of a synthetic mammography report with BI-RADS 3. See Multimedia Appendix 1 for an English translation. ACR: American College of Radiology. BIRADS: Breast Imaging Reporting and Data System.

MAMMOGRAPHIE IN ZWEI EBENEN VOM 11.10.2017.

Fragestellung/ Indikation:
Positive Familienanamnese (Schwester mit 62 Jahren). Tastbefund linke Mamma. Malignität?

Befund:
Keine Voruntersuchung.
Wenig fibroglanduläres Drüsengewebe beidseits.
Unauffällige Kutis und Subkutis beidseits.
In der linken Mamma umschriebener, runder Herdbefund, 19x12mm gross im unteren inneren Quadranten auf 8 Uhr. Keine Kalzifikation. Mamillendistanz 54mm. Kein weiterer Herdbedund in der linken Mamma. Keine Mikro- oder Makroverkalkungen.
Rechte Mamma ohne Herdbefund, Mikro- oder Makroverkalkung.
Keine suspekten Lymphknoten.

Beurteilung:
Kontrolle des Herdes links empfohlen. Sonst kein Anhalt für ein Malignom.
ACR Typ B beidseits
BIRADS 3 links
BIRADS 1 rechts.

Preprint JMIR Publications

To assess how closely the synthetic reports resemble the real reports in terms of their semantic content, we adopted a cosine similarity-based methodology: We use gte-Qwen2-1.5B-instruct as embedding model [37]. In detail, we sampled a stratified set of 21 real reports from DS.C. We then computed on the one hand the average similarity between this stratified set and the modelling dataset (DS.B) and on the other hand the average similarity between the 21 synthetic reports and DS.B. These three datasets are equally distributed across the BI-RADS categories (0-6). To improve comparability of this similarity-based approach, we additionally created a set of 21 German reports using ChatGPT (Free version used on 03/09/2024 with the prompt “Generate 21 radiology reports in German. They must be less than 700 words.") and also computed the average similarity between this LLM-generated dataset and DS.B.

Annotation process

The annotation process comprised three distinct phases: Initialization, quality improvement and final annotation and was designed to maximize the quality of extracted information in terms of clinical relevance, semantic and syntactic correctness. Below, we describe each phase in more detail.

The initialization phase was intended to define the use-case specific fact schema, describing the clinical information to be extracted from mammography reports, adhering to the underlying information model. Therefore, one computer scientist and one medical computer scientist (DR, JK) collaboratively created a first descriptive version of the fact schema based on the BI-RADS reporting guideline [8]. Moreover, a first set of annotation rules were defined and documented, inspired by existing literature [38]. Two clinicians (one radiologist, one radiation oncologist) then iteratively augmented the fact schema with information not mentioned in BI-RADS. This fact schema was checked regularly for adherence to the underlying information model. The initialization phase resulted in an initial fact schema and corresponding annotation guideline, documented in a web-based, versioned collaboration platform (Confluence, Atlassian). For this phase, DS.A was used.

The goal of the next phase, quality improvement, was to iteratively improve the fact schema and systematically revise the annotation guideline. Therefore, the coverage of the fact schema should be improved on the one hand, while on the other, Inter-Annotator-Agreement (IAA) should be maximized by augmenting and detailing the annotation guidelines. Quality improvement consisted of three iterations: In each iteration, two out of four clinicians applied the current fact schema and guideline version to annotate seven reports. The four clinicians included three physicians (two of them radiologists) and a medical student. After completion of the annotations, differences between clinicians were collaboratively discussed and resolved. Based on this discussion, the fact schema was adapted (addition, removal or joining of facts, anchors and modifiers) and the annotation guidelines were augmented, reducing ambiguities. Additionally, IAA (Krippendorff's alpha unitizing) was calculated on fact, anchor, and modifier levels for quantitative assessment. For this phase, DS.A was used.

After the quality improvement phase, both fact schema and annotation guideline were finalized and used to annotate DS.B by a total of three clinicians. For each report in DS.B, only the clinical findings section was annotated. After completion, generated annotations were automatically checked for syntactic adherence to the fact schema. Errors were resolved manually by the respective clinician.

As annotation tool, a local installation of the open-source annotation tool Inception was used [34]. During the project, the platform was migrated several times, spanning the versions 26.8 to 30.0. The tool provided all necessary key functionalities necessary to conduct the complete annotation process. The tool-specific annotation configuration file was generated automatically based on the fact schema by a proprietary tool. See Multimedia Appendix 2 for an example screenshot of the annotation interface.

Model development

As a basis for the fine-tuning process, we investigated four model variants , all based on the classical BERT architecture (110 million parameters): First, we applied medBERT.de, a BERT model further pre-trained on ~4.7 million German medical documents, including medical texts, clinical notes, research papers and other sources [39]. Second, we further pre-trained medBERT.de by performing masked language modelling on the pre-training corpus. We refer to the resulting model as InselBERT_multi, as the pre-training corpus contained reports related to multiple modalities. “Insel” refers to the name of the university hospital from which the pre-training data was retrieved. Third, we further pre-trained medBERT.de for three and ten epochs on a subset of the pre-training corpus, containing only mammography reports. We refer to the resulting models as InselBERT_mammo_03 (pre-trained for three epochs) and InselBERT_mammo_10 (pre-trained for ten epochs).

To extract clinical facts from mammography reports, a two-step model pipeline was implemented to meet the specific needs of extracting information according to the fact schema: First, clinical facts (continuous spans of text) are extracted based on fine-tuning the model variants for extractive question answering (QA model). Second, entities (anchors and modifiers) are extracted based on fine-tuning model variants for named entity recognition, implemented as sequence labelling task (NER model). For both models, the annotated modelling dataset (DS.B.ann) was programmatically transformed into the required input format for each of the two tasks. We fine-tuned each of the two pipeline models using medBERT.de, InselBERT_multi, and the two InselBERT_mammo variants.

Model deployment

After model development, the best performing fine-tuned QA and NER models (based on the

preprint JMIR Publications

averaged F1 score) were combined for inference using the open-source model serving framework BentoML [32]. Using BentoML, model binaries and associated files, serving logic and build configuration were packaged into a self-sustained unit (“Bento”), which was then containerized as a Docker image. This docker image was then deployed and exposed to an inference endpoint. This endpoint receives a mammography report and returns a structured JSON object containing the extracted facts and entities, including meta-information, e.g. label probabilities.

To integrate the model output into a clinical context and encode it according to established medical terminologies, we used the commercial software ID MACS® terminology server [31]. This approach enabled us to encode the extracted entities using the terminologies Wingert and SNOMED CT.

To evaluate the encoding procedure, we developed a custom, Java-based application (Integration server) that sends a mammography report to the inference endpoint, receives the extracted facts and contained entities and then sends each entity to the terminology server. The terminology server then returns the associated concepts from the specified terminologies.

Model evaluation

Our evaluation strategy comprises automated, quantitative evaluation according to state-of-the-art evaluation metrics (pre-trained model variants, fine-tuned QA model variants, fine-tuned NER model variants) as well as qualitative evaluation of the complete pipeline.

Further pre-training is evaluated by comparing model perplexity scores based on the validation split of the mammo corpus. Model perplexity is defined as “the level of perplexity when predicting the following symbol” [40] and computed as exponent of the averaged loss obtained from model validation.

The QA model variants are evaluated on a held-out validation split (15 %) applying the squad_v2 performance measure and bootstrapping [41]. According to the structure of the original squad_v2-dataset, the training split only consists of examples with exactly one answer (fact instance) for a question (fact type). Test- and validation set examples might have 1...n answers per question. The sequence labelling model variants are evaluated on a held-out validation split (10 %) applying the seq_eval performance measure [42].

In addition to the automated and separate evaluation of each developed model, a qualitative evaluation is carried out based on the synthetic report set. For each report, all facts and entities are extracted in a first step. To furthermore investigate the potential of our pipeline to support standardization, each extracted entity is sent to ID MACS® that returns concept candidates using two terminologies (SNOMED-CT and Wingert). All extracted entities and obtained terminology concepts are assessed by one radiation oncologist according to the following evaluation strategy: Each extracted fact and entity is marked as correct or incorrect. Each obtained concept is marked as plausible or not plausible. Furthermore, manual error analysis is conducted.

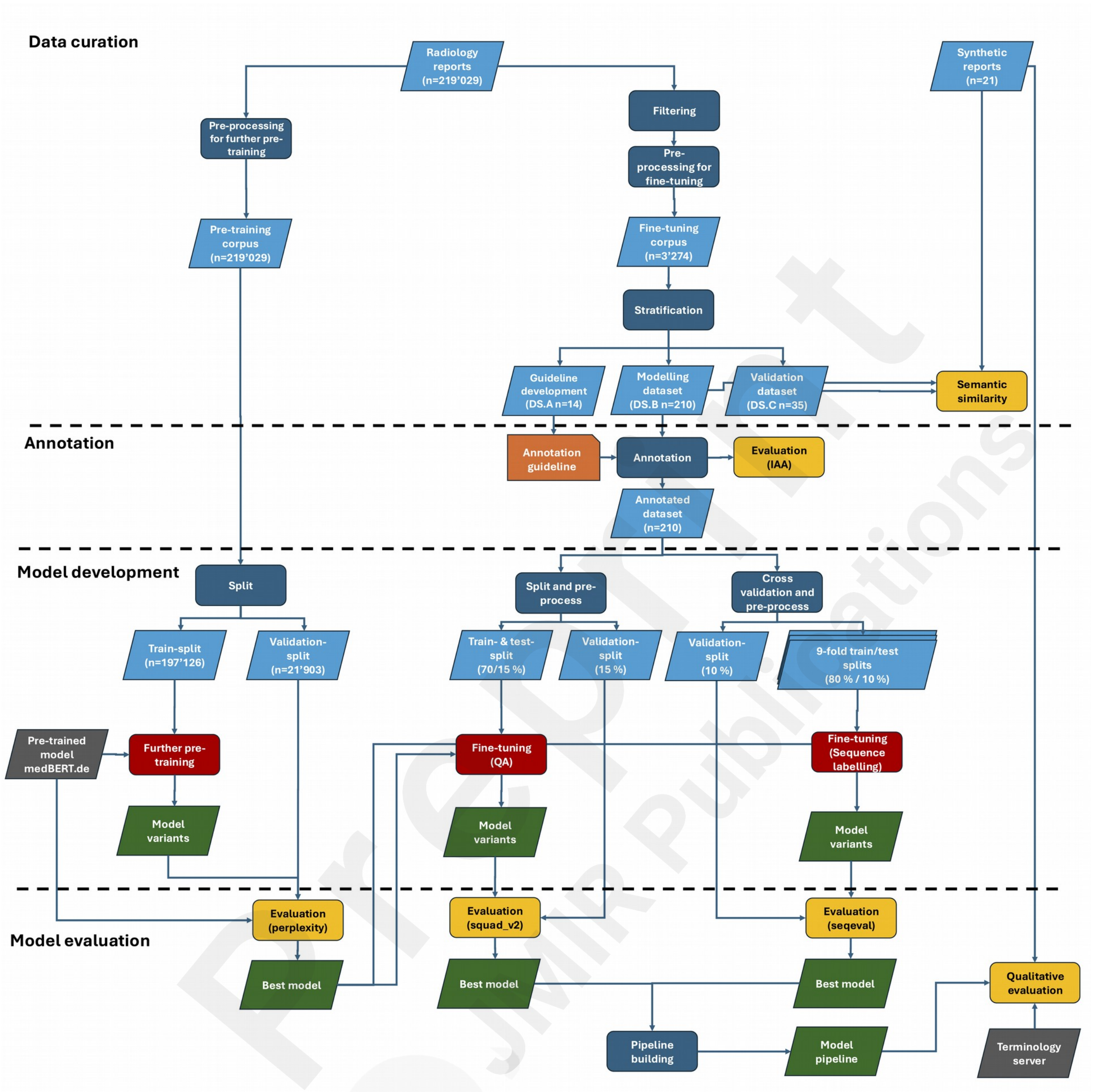
Results

Overview

The methodological steps described before can be grouped into the four distinct phases: data curation, annotation, model development, and model evaluation (see Figure 2). In the following subsections, we describe the results of the four phases.

Preprint JMIR Publications

Figure 2. Overview of model development



Preprint JMIR Publications

Data curation

The study, data acquisition and use were approved by the local ethics committee (Ethics committee Bern, BASEC 2022-01621). We used anonymized data covering the period of 2003-2023 provided by the university hospital of Bern (Inselspital Bern). The pre-training corpus comprised 219’029 radiology reports, obtained from Computer Tomography, Magnetic Resonance Imaging, X-Ray, and Sonography examinations. The mean number of tokens per report is 201, the median is 156, the standard deviation is 159, see Multimedia Appendix 3. For further pre-training, the corpus was split into a training (197’126 reports, 90 %) and a validation (21’903 reports, 10 %) set.

The fine-tuning corpus consisted of 3’274 mammography reports. The mean number of tokens per report is 158, the median is 146, the standard deviation is 63, see Multimedia Appendix 4. Encoding errors were fixed in 217 reports, 389 duplicate reports and 326 non-mammography reports were removed, resulting in a total of 2’559 reports before stratification.

The final datasets DS.A, DS.B and validation DS.C contained 14, 210 and 35 stratified reports, respectively. To allow batch inference of reports, a maximum document length of 512 tokens was defined, based on the fact that the 99.5 percentile of the report token length is 462 tokens, effectively ignoring eight reports during the sampling process.

Comparison of the synthetic dataset with DS.B shows only a minor decrease of the cosine similarity score as compared to the validation set (76,9 % and 79,1 %, respectively). In contrast, the comparison of the ChatGPT-generated reports with DS.B show a clear reduction of the cosine similarity score (57,3 %).

Annotation

The IAA results of the three rounds during the quality improvement phase are shown in Table 2. A total of 210 mammography reports (DS.B) was annotated according to the finalized fact schema and annotation guideline. For each report, only the clinical findings section was annotated. Therefore, out of 24 facts and 66 modifiers defined in the schema covering a complete mammography report, four facts (+ corresponding anchor) and 26 modifiers were not annotated, i.e. not present in the 210 annotated reports. Six facts (+corresponding anchor) and 14 modifiers that were annotated less than twenty times (defined as the minimum class frequency), were additionally excluded from the data. DS.B.ann comprised 14 fact types and 40 entity types (26 modifier types and 14 anchor types) for the domain of mammography, resulting in a total of 2’519 annotated fact instances. See Table S1 in Multimedia Appendix 5 for details regarding the frequencies of annotated facts and modifiers.

Table 2. IAA (Krippendorff’s alpha unitizing) after each annotation iteration of seven new reports.

	IAA Facts	IAA Anchors	Fact (= anchor) classes annotated (n)	IAA Modifiers	Modifier classes annotated (n)
Round 1					
	0.49	0.42	24	0.50	46
Round 2					
	0.74	0.64	22/ 21	0.60	45
Round 3					
	0.83	0.61	23/ 22	0.61	48

Model development

Masked language modelling was performed for three epochs applying a learning rate of 2e-5, a weight decay of 0.01 and a masking probability of 0.15.

For implementing the QA model, DS.B.ann was further separated into a train, test and validation split, containing 70 %, 15 % and 15 % of reports, respectively. Next, data was transformed into the squad_v2 format. This in turn resulted in 1’972, 274 and 273 examples per set, respectively. While the train and test set entries each contain only one question (fact type) with its associated answer in

Preprint JMIR Publications

the text (fact instance), the validation set entries might contain 1…n associated answers. This is always the case when one report contains multiple fact instances from the same fact type. For split generation, shuffling and seeding was applied.

Fine-tuning of the QA model was implemented based on an existing repository for extractive question answering [43]. Hyperparameters included a learning rate of 3e-5, 5 training epochs, maximum sequence length of 384 and a document stride of 128.

For the NER model, each annotated fact was transformed into an annotated training example, where each token of the fact was annotated according to the IOB format [44]. The 210 annotated reports contained 2’772 facts. 9-fold cross-validation was applied for model training, each fold comprising 80 % train data, and 10 % test data. The same validation split of 10 % of the data was shared among folds.

The PyTorch-based implementation of the NER model is based on the HuggingFace Transformers and Trainer API. Hyperparameters were defined as follows; a learning rate of 5e-5, a maximum of 100 training epochs, a weight decay of 1e-2, a batch size of 16 and the default AdamW optimizer. A seed was set manually. An early stopping strategy aborted the training process if model performance as measured on the test set dropped during five consequent epochs – in that case, the best model was saved, loaded, and final evaluation was performed on the held-out validation set.

Quantitative model evaluation

Further pre-training of model variants resulted in a reduction of the model perplexity score as shown in Table 3, obtained on an independent test set corresponding to 10 % of the available training data.

Table 3. Model perplexity scores before and after further pre-training.

	medBERT.de	InselBERT_multi	InselBERT_mammo(3 epochs)	InselBERT_mammo (10 epochs)
Model perplexity				
	1.21	1.12	1.11	1.09

Table 4 shows the results of fine-tuning each of the three further pre-trained model variants for extractive question answering including confidence intervals based on bootstrapping (599 samples), compared to the baseline model (medBERT.de). Table 5 shows the corresponding results of fine-tuning for sequence labelling, including standard deviation obtained by applying 9-fold cross-validation. Both tables show that InselBERT_mammo does not achieve significant improvement compared to the other models.

Table 4. Results of fine-tuning for extractive question answering including confidence interval and standard error (confidence level: 0.95, 599 resamples).

	medBERT.de	InselBERT_multi	**InselBERT_mammo (3 epochs)**	InselBERT_mammo (10 epochs)
Exact				

Preprint
JMIR Publications

match				
	79.49 (75.77-83.3), 1.84	78.02 (74.25-81.70), 1.91	**80.59** (77.33-83.99), 1.69	79.85 (76.33-83.41), 1.85
Averaged F1				
	90.09 (87.89-92.46), 1.16	89.78 (87.53-92.01), 1.13	**90.49** (88.36-92.47), 1.06	90.4 (88.28-92.53), 1.11

Table 5. Results of fine-tuning for sequence labelling averaged over all classes, based on 9-fold cross-validation(mean, standard deviation).

	medBERT.de	InselBERT_multi	**InselBERT_mammo (3 epochs)**	InselBERT_mammo (10 epochs)
Mean F1				
	0.81 (0.007)	0.81 (0.007)	0.81 (0.008)	0.81 (0.001)
Mean Recall				
	0.81 (0.008)	0.81 (0.01)	**0.82** (0.01)	0.81 (0.007)
Mean Precision				
	0.82 (0.009)	0.81 (0.007)	0.81 (0.01)	0.81 (0.009)
Mean Accuracy				
	0.82 (0.005)	0.82 (0.006)	0.82 (0.006)	0.82 (0.005)

Qualitative model evaluation

The final, dockerized model pipeline had a total size of 3.68 GB and was deployed on an institutional virtual machine without GPU (7.7 Gi RAM and 3.8 Gi Swap). Extracting all facts and entities from a single report took on average less than twenty seconds.

Next, this pipeline was used to extract facts and entities from every report in the synthetic dataset. This extraction resulted in a total of 205 extracted facts (14 types), of which 197 (96,10 %) were classified correctly. From 497 extracted entities (24 types), 495 (99,60 %) were classified correctly. Manual error analysis showed that misclassification of facts is limited to four fact types (“Mamillenregion beschrieben”, “Fremdmaterial beschrieben”, “Empfehlung für weitere Untersuchung”, “Asymmetrie beschrieben”) and misclassification of modifiers to a single modifier type (“Dignität”).

Regarding the automated normalization, suitable concepts from two terminologies (SNOMED-CT and Wingert) were retrieved for each extracted entity. Manual verification showed that out of a total of 3’582 concepts, 2’303 (64,29 %) were deemed plausible.

Preprint
© JMIR Publications

Discussion

Principal Results

In this paper, we present a novel, two-step IE pipeline based on the linguistic concept of frame semantics. First, extractive question answering is used to extract relevant text passages (“facts”) from free-text mammography reports. Next, for each fact, a sequence labelling model identifies a subset of attributes (“modifiers”), specified for each fact type.

Although further pre-training reduces model-complexity, it does not have a significant impact on neither of the two downstream-tasks (extractive QA and sequence labelling). Extractive QA reaches an averaged F1 score of > 90% and sequence labelling reaches a mean F1 score of > 80 %.

Using frame semantics in our extraction pipeline helps to organize radiological data in a way that aligns well with clinical practice: By defining frames for common scenarios, the model captures structured relationships between key terms (anchors) and their attributes (modifiers), improving clarity and consistency across reports. This approach makes it easier to add new fact types or modifiers without major adjustments, supporting standardized reporting that can scale effectively to different areas of radiology.

Compared to generative models, our approach has the following advantages: First, model output is directly linked to the original output on a token level, as both fine-tuned models perform sequence labelling on a token basis. Second, the underlying information model extracts relatively fine-grained information units based on modifiers: This detailed labelling facilitates further processing steps, e.g., normalization of terms. Last, both downstream tasks provide a certainty score for each labelled token, which could be used to dynamically filter out or highlight uncertain predictions.

Comparison to prior work

By using encoder-based model architectures, we ensure that the output of our pipeline is directly linked to the free-text in the report provided as input. In the case of model architectures partly or fully based on decoders (also known as generative models), this linkage cannot be directly ensured, but would have to be implemented manually (e.g., by prompting the model to provide the text span which it based its decision on and then verifying if this text span is contained in the original report).

Compared to the original IE pipeline proposed by Steinkamp et al., our implementation achieves a similar F1 score for fact extraction (Steinkamp: 91.0 % over 16 fact types, InselBERT_multi: 90.49 % over 14 fact types). According to their description of annotation frequencies, only two fact types (“radiologic finding was observed” and “anatomic region has property”) comprised 73.14 % of all fact annotations. Three fact types comprised less than 20 examples. These were still included in the final evaluation, whereas we decided to exclude all fact instances with less than twenty annotated examples. Unfortunately, performance measures on fact and modifier level are not reported by the authors. Comparability of both approaches is however limited, as a new set of facts was defined for our study.

While the original paper describing the squad_v2 dataset reported an F1 score of 66.3 % and an Exact Match score of 63.4 %, our approach seems to surpass these values by 24.2 %-points and 17.2 %-points, respectively (InselBERT_multi: 90.5 %, 80.6 %). Interestingly, our pipeline even exceeds human performance with a reported F1 score of 89.5 % by 0.99 %-points [41]. However, while the squad_v2 dataset contains various free-text questions, our model is restricted to the applied fact schema: Instead of free-text question, we train the model on fact names. This results in our model only being able to reliably extract the facts it has been trained on. Moreover, our dataset does not include unanswerable examples. In an unanswerable example, the provided context does not contain the answer to the posed question. Synthetically adding unanswerable examples to our dataset might positively impact model performance.

An important factor to be considered for real-world application of LLMs is model size, usually determined by number of parameters: While BERT has 340 million parameters, GPT-4, a recent decoder-only model, has an estimated amount of 1.8 trillion parameters, being approx. 5294 times larger [45]. Although model performance in general improves with model size, so does the required amount of training data and hardware resources: MedBERT.de, based on a German BERT variant, was pre-trained on a total of approx. 22 GB. GPT-3, the predecessor of GPT-4, however, was pre-trained on the CommonCrawl dataset of 45 TB of data [12]. According to Zhang et al., "pre-training PaLM [, a decoder-based model], requires around 2.5 × 10^{24} floating- point operations and takes 64 days when executed on 6,144 Google TPUv4 chips" [46]. Cloud-computing may alleviate these hardware requirements, although especially in healthcare, its adoption is complicated by existing data protection regulations. Moreover, as model architectures become larger, explainability of model outputs, defined as the "ability to explain or to present in understandable terms to a human"[47], decreases [20].

Due to the currently high demand for high performance computing hardware and especially GPUs, smaller models can be trained and inferred on relatively cheap and currently available hardware. This was the approach followed in this work.

Manual evaluation with synthetic data showed that the InselBERT pipeline was able to extract instances from ten out of 14 fact types and 23 out of 24 modifiers types with a precision of 100 % each. For the remaining four fact types and the single modifier types that included incorrect prediction, additional annotation might improve model performance. For the case study, prediction probabilities were not considered; increasing the prediction threshold for low-performing types might further increase class-level accuracy. Regarding the experimental normalization of extracted concepts, manual analysis shows that the low overall accuracy of 64,29 % is due to a large number of inplausible SNOMED-CT concepts. Performance might be increased by constraining subtype relationships for each entity (e.g., or by only using the Wingert terminology). We provide detailed results via OSF, see data availability statement.

Annotation process

We put a substantial effort into the generation of the fact schema and development of annotation guidelines. This was done to allow other research to reuse parts of our proposed fact schema, thereby saving time and resources. In future work, annotation effort might be reduced by applying active learning strategies as described by Jantscher et al [48].

As reported already by other researchers, the annotation process is the limiting factor in approaches like the one presented here. Challenges within the annotation process have already been described by Xia et al. and we had similar experiences [49]. Our annotation process presented several challenges, particularly in the communication between informaticians and clinicians. The project encountered issues such as frequent changes in team members and unclear instructions. Future strategies include the generation of silver-standard labels, with clinicians focusing solely on correction. Additionally, it is important to note the limited availability of clinicians' time when designing and implementing annotation projects [50].

From the annotators' perspective, while numerous components of the radiology reports proved to be formalizable, certain challenges persisted: Structurally, issues such as determining referentiality between phrases and the modification relationships of adjectives presented difficulties. Additionally, variability in report-writing conventions across individual clinicians posed considerable obstacles. This variability was particularly problematic given our effort to base annotation guidelines on the BI-RADS criteria for mammographies, resulting in difficulties when annotating reports that did not adhere to these standards.

Nonetheless, the establishment of detailed annotation guidelines enabled thorough annotation of the majority of the clinical findings sections within the radiological reports. We want to emphasize the need to implement an extensive trial phase for clinical annotation projects. This phase should involve (1) the initial development of guidelines by technicians as well as clinicians and (2) subsequent, iterative development and testing of these guidelines. We saw that, after the third iteration, an IAA of >0.7 was effective in resolving ambiguities in the annotation guidelines.

In future studies, it is recommended to separate a guideline development team from an annotation team. Additionally, employing a larger set of reports (<10) for each iteration is recommended, as increasing the sample size will enhance the generalizability of the guidelines instead of repeating to create specialized rules for a smaller set of reports.

Limitations

In the following, we point out the limitations of this study to be considered. One major prerequisite of the applied information model is that all information related to a fact is located within one unbroken text span. Therefore, information mentioned outside of facts can currently not be considered. Specifically for the use-case of mammography, we noticed that clinicians report findings for each breast separately (Left breast: <paragraph>. Right breast: <paragraph>). In future, we want to investigate how to include this external information in the frame semantics approach.

Moreover, due to time constraints during the annotation process, it was decided to only annotate the results sections of the included reports. There might be relevant information in the other sections of the mammography report. To speed up the annotation process, our underlying framework is centered around the idea of re-using existing facts and annotations for other clinical use-cases; whether this approach is realistic is subject to an ongoing follow-up project.

Regarding model development, we highlight that our pipeline comprising two different models was not trained end-to-end. Instead, we developed and evaluated each of the two models separately, before carrying out end-to-end human evaluation using synthetic reports only. Currently, our model is only capable of analyzing German mammography reports. However, using annotation-preserving translation of our training data, our pipeline could be easily adapted to any other language. This task is also known as annotation projection and has been described in the literature [51,52].

Unfortunately, our core dataset cannot be shared due to institutional restrictions: However, it is possible to evaluate our models based on the published synthetic dataset, which should resemble the original reports very well. Furthermore, our pipeline is trained on data obtained by only one institution. We therefore cannot assess the adaptability of our pipeline on external data and generalizability may be limited.

One major limitation regarding the implementation of extractive question answering is the fact that we did not generate unanswerable questions based on our training data. The original dataset published with the metric used in this paper, squad (version 2), contains approx. 30 % of unanswerable questions. We hypothesize that programmatically creating and adding unanswerable entries, performance should further increase. Furthermore, we point out that both downstream tasks are evaluated differently: While extractive QA was developed using a fixed train/test/validation split and evaluated using bootstrapping, the sequence labelling model was developed using 9-fold cross-validation and a shared validation set.

Future work

We note that there are several open research questions related to our proposed approach that needs further investigation and might lead to performance increase:
Currently, only the complete text-span annotated as a fact is used for training the sequence labelling model. We hypothesize that the model would benefit from a bigger context window, including either the whole report or additional leading and trailing tokens surrounding the text span, as implemented by Steinkamp et al. [27]. Another strategy to improve sequence labelling performance is concatenating the model input with encoded information on the specific fact type. A similar approach was described by Kuling et al. who encoded and concatenated the corresponding report section of each sentence to be labelled [53].

Another open aspect relates to the sharing of modifiers between fact types as well as experimenting with different modifier aggregation levels: For example, all modifiers describing an anatomical location might be aggregated and/or shared between all facts. It also remains open whether our QA model is capable of inferring new fact types and whether providing a textual description of the fact type instead of its title might improve performance.

Further pre-training of model variants showed little to no effect on downstream task performance. However, as a base model, we applied Medbert.de - a model already pre-trained on > 4.7 million German medical documents, comprising approx. 3.6 million radiology reports. We hypothesize that our training data set was simply too small to impact the pre-trained model parameters. Increasing further pre-training epochs, changing hyperparameters like learning rate and/or training a model from scratch solely based on our training data set are potential strategies for improving model performance that we have not yet investigated.

Last, we chose an encoder-based approach due to the inherent interpretability of model outcomes as compared to generative models that by design cannot perform token-level predictions. However, recent research shows that including diagnostic reasoning in generative models might improve the assessment of outcomes for clinicians [54]. Due to the high performance and smaller sizes of recent generative models, applying frame semantics to generative models might offer additional advantages.

Conclusions

In this paper, we demonstrated the feasibility of integrating frame semantics with a BERT-based architecture for extracting information from mammography reports with the purpose of automatic structuring. Our approach allows clinicians to specify the aspects of relevance to be extracted from the reports. Additionally, using encoder-based model architectures, the output of our pipeline is directly linked to the free-text in the report provided as input. This supports the explainability of the extraction results which is crucial in settings where it has to be ensured that the system output is correct. We conclude that even with a small annotated dataset and a rather small language model, the quality of IE is still good and comparable to human extraction quality.
In future work, we will validate the results and test the application of the approach to reports from other radiology departments. Additionally, the transfer to other examinations and report types will be

studied. The suggested approach was designed in a way that information to be extracted can be defined by clinicians. Together with a dataset annotated according to a new fact schema, the approach could be transferred easily.

Acknowledgements

We thank all clinicians from the university hospital of Bern for their valuable contributions during the annotation process. We thank Dr. Harald Bonel for contributing to generation of the synthetic dataset. Grazia Maria Cereghetti obtained the ethics approval.

CRediT author contributions

Reichenpfader: Methodology, Software, Visualization, Investigation, Writing – Original Draft, Writing - Review & Editing
Knupp: Methodology, Software
Von Däniken: Resources, Writing – Review & Editing
Gaio: Software, Writing – Original Draft
Dennstädt: Resources, Writing – Review & Editing
Cereghetti: Resources, Data Curation
Sander: Conceptualization
Hiltbrunner: Writing – Original Draft, Resources
Nairz: Conceptualization, Project administration, Writing – Original Draft
Denecke: Conceptualization, Supervision, Writing - Review & Editing

All authors reviewed and approved the manuscript for publication.

Conflicts of Interest

André Sander is member of the executive board at ID Suisse AG.

Abbreviations

ACR: American College of Radiology
API: Application Programming Interface
BERT: Bidirectional Encoder Representations from Transformers
BI-RADS: Breast Imaging-Reporting and Data System
CAS: Common Analysis Structure
CDE: Common Data Element
(E)QA: (Extractive) question answering
ESR: European Society of Radiology
FMH: Swiss Medical Association
IAA: Inter-annotator-agreement
IE: Information extraction
LLM: Large language model
NER: Named Entity Recognition
NLP: Natural Language Processing
QA: Question answering
RIS: Radiology information system
UI: User interface

preprint JMIR Publications

Data Availability

We make the synthetic dataset, the original annotation guideline in German, all source code, and detailed evaluation results publicly available via OSF (https://doi.org/10.17605/OSF.IO/C7GFE) and Zenodo (added after acceptance).

References

1. Pinto dos Santos D, Kotter E, Mildenberger P, Martí-Bonmatí L, European Society of Radiology (ESR). ESR paper on structured reporting in radiology—update 2023. Insights Imaging 2023 Nov 23;14(1):199. doi: 10.1186/s13244-023-01560-0

2. Langlotz CP. Radiology report: a guide to thoughtful communication for radiologists and other medical professionals. San Bernardino, CA: CreateSpace Independent Publishing Platform; 2016. ISBN:978-1-5151-7408-0

3. Ganeshan D, Duong P-AT, Probyn L, Lenchik L, McArthur TA, Retrouvey M, Ghobadi EH, Desouches SL, Pastel D, Francis IR. Structured Reporting in Radiology. Academic Radiology 2018 Jan;25(1):66–73. doi: 10.1016/j.acra.2017.08.005

4. Pinto Dos Santos D, Hempel J-M, Mildenberger P, Klöckner R, Persigehl T. Structured Reporting in Clinical Routine. Fortschr Röntgenstr 2019 Jan;191(01):33–39. doi: 10.1055/a-0636-3851

5. Percha B, Zhang Y, Bozkurt S, Rubin D, Altman RB, Langlotz CP. Expanding a radiology lexicon using contextual patterns in radiology reports. Journal of the American Medical Informatics Association 2018 Jun 1;25(6):679–685. doi: 10.1093/jamia/ocx152

6. Martin-Carreras T, Cook TS, Kahn CE. Readability of radiology reports: implications for patient-centered care. Clinical Imaging 2019 Mar;54:116–120. doi: 10.1016/j.clinimag.2018.12.006

7. Olthof AW, Leusveld ALM, De Groot JC, Callenbach PMC, Van Ooijen PMA. Contextual Structured Reporting in Radiology: Implementation and Long-Term Evaluation in Improving the Communication of Critical Findings. J Med Syst 2020 Sep;44(9):148. doi: 10.1007/s10916-020-01609-3

8. American College of Radiology. ACR BI-RADS atlas: breast imaging reporting and data system. 5th ed. Reston, Virginia; 2013.

9. Magny SJ, Shikhman R, Keppke AL. Breast Imaging Reporting and Data System. StatPearls Treasure Island (FL): StatPearls Publishing; 2024. PMID:29083600

10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is All you Need. Advances in Neural Information Processing Systems Curran Associates, Inc.; 2017.

11. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. Available from: http://arxiv.org/abs/1810.04805 [accessed Jan 17, 2023]

Preprint JMIR Publications

12. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 1877–1901.

13. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv; 2023. Available from: http://arxiv.org/abs/1910.10683 [accessed May 5, 2024]

14. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NG, Löffler CML, Schwarzkopf S-C, Unger M, Veldhuizen GP, Wagner SJ, Kather JN. The future landscape of large language models in medicine. Commun Med Nature Publishing Group; 2023 Oct 10;3(1):1–8. doi: 10.1038/s43856-023-00370-1

15. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med Nature Publishing Group; 2023 Aug;29(8):1930–1940. doi: 10.1038/s41591-023-02448-8

16. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med Nature Publishing Group; 2024 Feb 27;1–9. doi: 10.1038/s41591-024-02855-5

17. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. Nature Nature Publishing Group; 2023 Aug;620(7972):172–180. doi: 10.1038/s41586-023-06291-2

18. Choi HS, Song JY, Shin KH, Chang JH, Jang B-S. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J The Korean Society for Radiation Oncology; 2023 Sep 21;41(3):209–216. doi: 10.3857/roj.2023.00633

19. Hu D, Liu B, Zhu X, Lu X, Wu N. Zero-shot information extraction from radiological reports using ChatGPT. International Journal of Medical Informatics 2024 Mar 1;183:105321. doi: 10.1016/j.ijmedinf.2023.105321

20. Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, Wang S, Yin D, Du M. Explainability for Large Language Models: A Survey. ACM Trans Intell Syst Technol 2024 Apr 30;15(2):1–38. doi: 10.1145/3639372

21. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature Nature Publishing Group; 2024 Jun;630(8017):625–630. doi: 10.1038/s41586-024-07421-0

Preprint JMIR Publications

22. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6:120. PMID:37414860

23. Reichenpfader D, Müller H, Denecke K. Large language model-based information extraction from free-text radiology reports: a scoping review protocol. BMJ Open British Medical Journal Publishing Group; 2023 Dec 1;13(12):e076865. PMID:38070902

24. Gorenstein L, Konen E, Green M, Klang E. Bidirectional Encoder Representations from Transformers in Radiology: A Systematic Review of Natural Language Processing Applications. Journal of the American College of Radiology Elsevier; 2024 Jun 1;21(6):914–941. PMID:38302036

25. Saha A, Burns L, Kulkarni AM. A scoping review of natural language processing of radiology reports in breast cancer. Front Oncol 2023;13:1160167. PMID:37124523

26. Gholipour M, Khajouei R, Amiri P, Hajesmaeel Gohari S, Ahmadian L. Extracting cancer concepts from clinical notes using natural language processing: a systematic review. BMC Bioinformatics 2023 Oct 29;24(1):405. PMID:37898795

27. Steinkamp JM, Chambers C, Lalevic D, Zafar HM, Cook TS. Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning. J Digit Imaging 2019 Aug;32(4):554–564. PMID:31218554

28. Fillmore CJ. Chapter 10 frame semantics. In: Geeraerts D, editor. Cognitive linguistics: Basic readings Berlin, New York: De Gruyter Mouton; 2006. p. 373–400. doi: doi:10.1515/9783110199901.373ISBN:978-3-11-019990-1

29. Reichenpfader D, Knupp J, Sander A, Denecke K. RadEx: A Framework for Structured Information Extraction from Radiology Reports based on Large Language Models. arXiv; 2024. doi: 10.48550/arXiv.2406.15465

30. Confluence | Your Remote-Friendly Team Workspace | Atlassian. Available from: https://www.atlassian.com/software/confluence?&aceid=&adposition=&adgroup=144569217489&campaign=18443387038&creative=666242883243&device=c&keyword=atlassian%20confluence&matchtype=e&network=g&placement=&ds_kids=p73335758169&ds_e=GOOGLE&ds_eid=700000001542923&ds_e1=GOOGLE&gad_source=1&gbraid=0AAAAAD2bkRdXosciijz2uW0VoXu-xEF1A&gclid=CjwKCAjwyo60BhBiEiwAHmVLJdjcfuSRjwL8Ob4cyvvlDjq1yghCBSOWxML4T65oF-9HzKF6fx8qIhoC8ZEQAvD_BwE&gclsrc=aw.ds [accessed Jul 2, 2024]

31. ID Berlin. ID MACS® | Terminology server. 2024. Available from: https://www.id-berlin.de/produkte/nlp-forschung/id-macs/ [accessed Mar 21, 2024]

32. Yang C, Sheng S, Pham A, Zhao S, Lee S, Jiang B, Dong F, Guan X, Ming F. BentoML: The framework for building reliable, scalable and cost-efficient AI application. 2024. Available from: https://github.com/bentoml/bentoml [accessed Mar 21, 2024]

33. GitHub Packages: Your packages, at home with their code. GitHub. 2024. Available from: https://github.com/features/packages [accessed Jul 2, 2024]

preprint JMIR Publications

34. de Castilho RE, Klie J-C, Kumar N, Boullosa B, Gurevych I. INCEpTION - corpus-based data science from scratch. Lisbon, Portugal; 2018. Available from: http://tubiblio.ulb.tu-darmstadt.de/106982/

35. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Scao TL, Gugger S, Drame M, Lhoest Q, Rush AM. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv; 2020. doi: 10.48550/arXiv.1910.03771

36. Honnibal M, Montani I, Van Landeghem S, Boyd A. spaCy: Industrial-strength Natural Language Processing in Python. 2020. doi: 10.5281/zenodo.1212303

37. Li Z, Zhang X, Zhang Y, Long D, Xie P, Zhang M. Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:230803281 2023;

38. Liu S, Fu S, Liu H. Informatics playbook, Chapter 9: Best practices of annotating clinical texts for information extraction tasks¶. The National Center for Data to Health (CD2H); 2022. Available from: https://playbook.cd2h.org/en/latest/chapters/chapter_9.html

39. Bressem KK, Papaioannou J-M, Grundmann P, Borchert F, Adams LC, Liu L, Busch F, Xu L, Loyen JP, Niehues SM, Augustin M, Grosser L, Makowski MR, Aerts HJ, Löser A. MEDBERT.de: A Comprehensive German BERT Model for the Medical Domain. Expert Systems with Applications 2023 Mar;237:121598. doi: 10.1016/j.eswa.2023.121598

40. Huyen C. Evaluation metrics for language modeling. The Gradient 2019; Available from: https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

41. Rajpurkar P, Jia R, Liang P. Know What You Don't Know: Unanswerable Questions for SQuAD. arXiv; 2018. Available from: http://arxiv.org/abs/1806.03822 [accessed Jun 26, 2024]

42. Nakayama H. Seqeval: A Python framework for sequence labeling evaluation. 2018. Available from: https://github.com/chakki-works/seqeval

43. Debut L. huggingface/transformers: Question answering. HuggingFace; 2024. Available from: https://github.com/huggingface/transformers/blob/main/examples/pytorch/question-answering/README.md [accessed Jun 3, 2024]

44. Tjong Kim Sang EF, Buchholz S. Introduction to the CoNLL-2000 Shared Task' Chunking. 2000; Available from: https://aclanthology.org/W00-0726

45. Patel D. GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE. 2023. Available from: https://www.semianalysis.com/p/gpt-4-architecture-infrastructure [accessed Nov 5, 2024]

46. Zhang L, Liu X, Li Z, Pan X, Dong P, Fan R, Guo R, Wang X, Luo Q, Shi S, Chu X. Dissecting the Runtime Performance of the Training, Fine-tuning, and Inference of Large Language Models. arXiv; 2023. Available from: http://arxiv.org/abs/2311.03687 [accessed Apr 10, 2024]

47. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. arXiv;

Preprint
JMIR Publications

2017. Available from: http://arxiv.org/abs/1702.08608 [accessed Apr 10, 2024]

48. Jantscher M, Gunzer F, Kern R, Hassler E, Tschauner S, Reishofer G. Information extraction from German radiological reports for general clinical text and language understanding. Sci Rep Nature Publishing Group; 2023 Feb 9;13(1):2353. doi: 10.1038/s41598-023-29323-3

49. Xia F, Yetisgen-Yildiz M. Clinical Corpus Annotation: Challenges and Strategies. Proceedings of the third workshop on building and evaluating resources for biomedical text mining (BioTxtM'2012) in conjunction with the international conference on language resources and evaluation (LREC) istan; 2012.

50. Wei Q, Franklin A, Cohen T, Xu H. Clinical text annotation – what factors are associated with the cost of time? AMIA Annual Symposium Proceedings 2018 Dec 5;2018:1552. PMID:30815201

51. García-Ferrero I, Agerri R, Rigau G. T-Projection: High Quality Annotation Projection for Sequence Labeling Tasks. In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023 Singapore: Association for Computational Linguistics; 2023. p. 15203–15217. doi: 10.18653/v1/2023.findings-emnlp.1015

52. Akbik A, Vollgraf R. ZAP: An Open-Source Multilingual Annotation Projection Framework. In: Calzolari N, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, Tokunaga T, editors. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) Miyazaki, Japan: European Language Resources Association (ELRA); 2018. Available from: https://aclanthology.org/L18-1344 [accessed Jun 26, 2024]

53. Kuling G, Curpen B, Martel AL. BI-RADS BERT and Using Section Segmentation to Understand Radiology Reports. J Imaging 2022;8(131). doi: 10.3390/jimaging8050131

54. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. npj Digit Med Nature Publishing Group; 2024 Jan 24;7(1):1–7. doi: 10.1038/s41746-024-01010-1

Preprint JMIR Publications

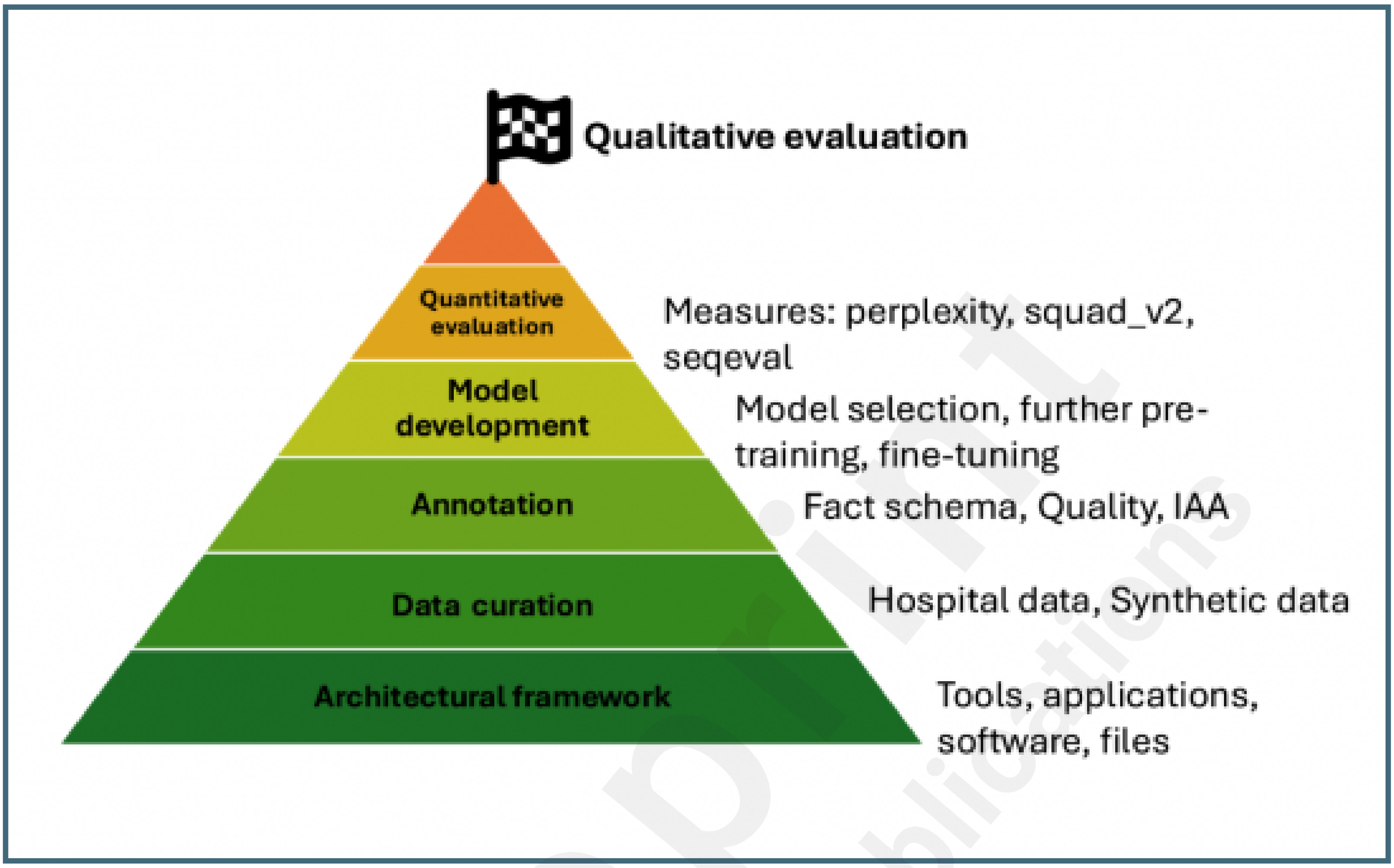
Supplementary Files

Preprint
JMIR Publications

Figures

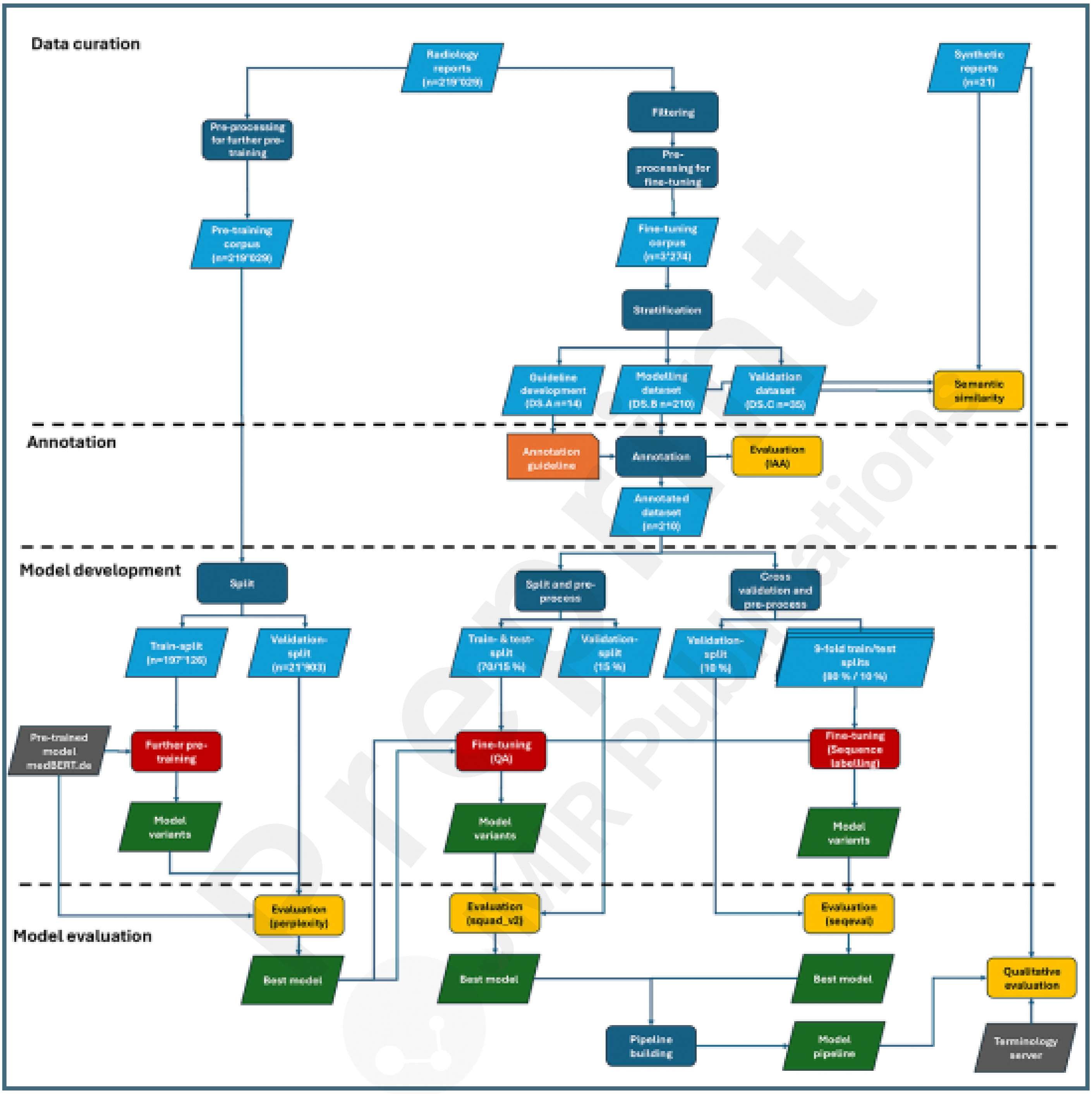
Preprint
JMIR Publications

Methodology overview. IAA: Inter-annotator-agreement.


Qualitative evaluation
Quantitative evaluation
Measures: perplexity, squad_v2, seqeval
Model development
Model selection, further pre-training, fine-tuning
Annotation
Fact schema, Quality, IAA
Data curation
Hospital data, Synthetic data
Architectural framework
Tools, applications, software, files

Preprint JMIR Publications

Overview of model development.



Preprint
JMIR Publications

Multimedia Appendixes

Preprint JMIR Publications

Translation of example report.
URL: http://asset.jmir.pub/assets/929a68d94cf660c1f1527ba994a32d1e.docx

Example of an annotated report (screenshot from the applied annotation tool Inception). The example comprises five annotated sentences within a single report. Facts are annotated in red, anchor entities in green and modifiers in blue. For overlapping facts, each modifier is additionally linked to all of its corresponding anchor entities as seen in line five. Translation: Dense gland parenchyma. Regular visualization of the cutis, subcutis and nipple region. No suspicious, spiculated focal findings. No suspicious, linear, branched or pleomorphic microcalcifications. No suspicious lymph nodes as far as recorded.
URL: http://asset.jmir.pub/assets/67b500a4551749b0bda87b4fd662c4c0.png

Token characterstics of pre-training data.
URL: http://asset.jmir.pub/assets/9e2389701ee32ea7566bc656baf8a14e.png

Token statistics of fine-tuning data (only mammographies).
URL: http://asset.jmir.pub/assets/768ab2dbc2de0d886f6ed1d0c2f294db.png

Number of annotated fact and modifier instances.
URL: http://asset.jmir.pub/assets/4b36c27a3636aac3d496f2843000974e.docx

Preprint JMIR Publications