Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Natural Language Processing Techniques for Extracting Diagnostic Information from Radiology Reports

Dr B. Ranganatha Rao[1*], Smt. K. L. Madhuri[2], D. Snehalatha[3]

[1*]Assistant Professor, Osmania University Hyderabad, Email: ranganath.ballarapu@gmail.com
[2]Associate Professor, Nanakram Bhagwandas Science College, Email: klmadhuri24@gmail.com
[3]Assistant professor, N.B.Science College, Email: snehalatha10@gmail.com

***Corresponding Author:** Dr B. Ranganatha Rao
*Assistant Professor, Osmania University Hyderabad, Email: ranganath.ballarapu@gmail.com

ARTICLE INFO

Received: 13 Nov 2023

Accepted: 24 Jan 2024

ABSTRACT

The ability of extracting diagnostic information using Natural Language Processing (NLP) in unstructured radiology texts is becoming of increasing importance, and it is necessary due to the rapid growth of imaging studies. This paper dwells on automating the process of extracting diagnostic labels on English chest X-rays report texts by applying an NLP pipeline that comprises a hybrid of rule-based (chexpert, NegBio) and transformer-based (BioBERT, ClinicalBERT) components. This required the training and deployment of a machine-learning model process to parse radiology reports associated with a large-scale dataset, MIMIC-CXR-JPG and over 14 key pathologies and No Finding were attributed; with DICOM metadata also incorporated and structured to support downstream analysis with robustness. NLP systems showed high accuracy in prevalent findings like cardiomegaly, edema and pleural effusion (F1-scores >0.85), but displays reduced accuracy in the low frequency or linguistically confusing classes, and there is a severe class imbalance present: most findings are classified as No Finding. Exploratory data analysis and benchmarking indicated that rule-based tools are efficient when deployed on common patterns, whereas transformer models utilized to a larger extent contextual and uncertain information and required more substantial computational resources. The researchers summarize that hybrid NLP methods have the potential to automate and standardise the extraction of labels in reports of chest radiology, thus facilitating clinical records and enabling large-scale studies, as well as the improvement of the basis of AI-based diagnosis systems, although future research is required to solve the issue of generalisability and challenging linguistic constructions.

Keywords: Natural Language Processing, Chest X-ray, Radiology Reports, Automated Label Extraction, Clinical NLP

1. INTRODUCTION

Radiology report is a clinical report that documents the interpretation of diagnostic tests like chest X-rays, CT scans, and MRI. These reports are usually written in free-text form and they include important information about anatomical abnormalities, pathological results, and clinical impressions. As the number of imaging studies produced in hospitals and diagnostic facilities around the world grows, there is an urgent requirement to have automated systems that can mine information that is pertinent to the diagnosis of a patient out of these unstructured narrative reports. NLP is a sub-field of artificial intelligence, which offers the computational basis to convert these free-text radiology reports into structured data that can be indexed, queried, and analyzed to be used clinically and in research (Johnson et al., 2019).

Over the last ten years, the opportunities of NLP in radiology have increased by orders of magnitude with the advent of large-scale, de-identified datasets such as MIMIC-CXR-JPG, which contains over 370,000 labeled chest radiographs and their associated reports (Johnson et al., 2019). Such a corpus has been used to develop

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

rule-based tools, including CheXpert, which utilizes regular expressions and linguistic heuristics to label 14 of the most frequently occurring radiological findings, such as cardiomegaly, pneumonia, and pleural effusion (Irvin et al., 2019). At the same time, such systems as NegBio and MetaMap Lite were implemented to detect negation and uncertainty in narratives of radiology reports, which is one of the oldest and most notorious problems in clinical NLP: how to interpret some statements like the absence of pneumothorax or the possibility of effusion (Peng et al., 2018; Demner-Fushman et al., 2017).

Nevertheless, these contributions have limitations in the sense that the current techniques do not always work well with complex or non-standardly structured reports. A significant number of rule-based methods are interpretable and computationally efficient but are not resilient to changes in syntax, grammar, and radiologist reporting styles. They are susceptible, especially when it comes to errors in the detection of negation, and uncertainty treatment (Peng et al., 2018). Deep bidirectional Transformer models have shown much improvement in the contextual understanding and generalization by using deep bidirectional embeddings that are trained on large biomedical corpora (Alsentzer et al., 2019; Zhang et al., 2019). The models have proven to be on the cutting edge of performance when it comes to tasks such as named entity recognition (NER), relation extraction, and multi-label classification. Nonetheless, their use can be demanding in terms of computational resources and annotated datasets, which cannot be readily available in a clinical environment.

These difficulties are further worsened by the fact that distribution of findings on chest radiology reports is inherently unbalanced. To illustrate, most of the MIMIC-CXR studies have been classified as No Finding with the less common diagnoses such as lung lesions and fractures having fewer than 2% of cases (Irvin et al., 2019). This skewness is a challenge to both rule-based and neural approaches, especially when they have to achieve high recall without compromising on precision. In addition, newer methods like RadGraph and Cliner 2.0 increase the accuracy of extraction by modeling not only the recognition of entities but also the association between the clinical concepts but are still developing and require additional empirical validation (Boag et al., 2018; Banerjee et al., 2018).

A second limitation is that there is no generalizability between imaging modalities and institutions. The majority of systems have been trained on English radiology reports of a single dataset i.e. MIMIC-CXR or CheXpert. Such systems might not serve the same well with the data provided by other hospitals or reports that are in other languages or forms. Although large-scale language models have demonstrated potential to adapt to a variety of contexts, the addition of fine-tuned layers of negation detection and pathology-specific knowledge bases is a key aspect of attaining a clinically effective performance (Su et al., 2025). Despite the fact that there is some more recent work trying to add multi-task learning and reinforcement-based optimization to enhance label fidelity and robustness, these approaches are still experimental (Ke et al., 2025).

This paper is devoted in particular to the extraction of diagnostic labels in English-language chest X-ray radiology reports through both rule-based and transformer-based NLP methods. It employs MIMIC-CXR-JPG dataset which is a comprehensive and publicly available benchmark to check. It is narrow in its focus to the textual report parsing and does not use the visual data of X-ray images. Also, the extraction of procedural or therapeutic information is not discussed in the study and the focus is on the correct retrieval of the findings of diagnosis and the modifiers of negation or uncertainty.

The bigger picture of this study is to improve automation and standardization of clinical documentation in radiology. This study will attempt to systematically integrate and assess the performance of rule-based models, such as CheXpert and NegBio, with neural models, such as Clinical BERT, to determine the performance shortcomings, to measure the accuracy of label extraction per pathology and suggest improvements of current NLP pipelines. This is especially relevant as the use of artificial intelligence in radiological practice continues to grow, and structured data based on free-text reports can enhance decision support, efficiency of radiologists, and clinical research. Moreover, the results will support the creation of high-quality training labels to the deep learning models that are based on the structured annotations of unstructured clinical reports.

This study has clinical and academic implications. It is part of an accumulating literature on medical text mining, it adds evidence to the practicality of hybrid NLP systems, and it offers empirical evidence of the performance of available tools on real-world radiology data. The results may facilitate the creation of hospital

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

wide information retrieval systems, automated disease surveillance systems and even radiology education and audit tools.

In order to meet the goals of the research, the following aims were formulated:

- To develop and experiment with a hybrid NLP pipeline containing rule-based and transformer-based methods of radiological diagnosis extraction in chest X-ray reports.
- To evaluate the efficiency of the systems of negation and uncertainty detection (e.g., NegBio, MetaMap, ClinicalBERT) under different pathologies.
- To determine the limitations of current NLP systems in generalizing to a variety of report types and suggest solutions to improve them so that they can be more widely applied in clinical settings.

Overall, the current study aims to further the automated interpretation of chest radiology reports based on a well-defined and clinically driven natural language processing pipeline.

2. LITERATURE REVIEW

The use of Natural Language Processing (NLP) has emerged as a root technology in the automated interpretation of radiology reports, and it is possible to imagine converting the unstructured free-text accounts into structured and clinically actionable information. With digital medical imaging growing in scale, there is growing demand to automate radiology report interpretation, hence a wide range of research studies has ensued which utilize entity recognition, relation extraction, uncertainty detection and multimodal learning. Upon examination of the available literature, a pattern is emerged in that contributors have since transformed their rule-based systems to the current transformer-based and hybrid approaches to neural networks which are much more flexible, scalable and relevant to clinical practice.

The construction of large data sets of chest radiographs, including ChestX-ray8 (Wang et al., 2017), can be considered one of the first and largest contributions in this direction. This dataset, which involves over 100 000 frontal-view X-ray images, allowed the researchers to explore the possibilities of weakly supervised classification and localization of thoracic diseases and do so using only the image-level labels that were extracted out of free-text reports. Researchers proposed a recurrent neural cascade model (Shin et al., 2016) which utilized report-level annotation to generate disease labels automatically given a chest X-ray image and the model effectively showed that even shallow annotation can vastly increase results of deep learning models trained on radiological images.

When NLP got matured, efforts were made in the direction of extraction of organized entities and the semantic relations of these entities in radiology reports. Among the most progressive attempts in this direction is the RadGraph model that was proposed by Jain et al. (2021). RadGraph not only recognizes clinical entities (e.g., the anatomical structures and observations), but also the relations among them (e.g., the presence of, located at), and thus allows get very specific information out of the radiology narratives. This entity-relation model backs clinical decision support systems, automated audit, and the medical knowledge graph development. On the same note, Yoon et al. (2019) introduced Collabonet, or a collaborative system of deep neural networks, improving dramatically on biomedical named entity recognition (NER) through a combination of a number of nitro-terminated sub-models that interacted with one another during both training and inference stages.

It has changed how NLP is used in biomedical applications due to transformer-based models. BioBERT, a pre-trained language model that focuses on biomedical corpora, was one of them and achieved the state-of-the-art performances in a diverse range of biomedical NLP tasks, including NER, question answering, and relation extraction (Lee et al., 2020). The RoBERTa, which Liu et al. (2019) implemented as another transformer model, boosted the performance of the BERT because it maintained more aggressive pretraining techniques, and has since been transferred into clinical fields. Such models have substantial advantage of the rule-based approaches, particularly with regard to cross-institutional and cross-reporting-style generalization.

Another direction in which the integration of NLP is studied is the integration with the ontology-based systems. Tiaiba et al. (2025) have carried out an extensive survey of the synergy between ontologies and machine

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

learning models in the health informatics. They worked on highlighting the importance of learning standardization on terms based on ontology in enhancing consistency and accuracy of the information extraction on clinical narratives. The SNOMED CT and RadLex ontologies give standard vocabularies that help diminish uncertainty and lead to cross-institutional interoperability, which enhances the dependability of NLP pipelines in radiology.

The paper by Hassanpour and Langlotz (2016) focused on discussing the issue of multi-institutional variability in the report writing in radiology. They suggested an information-extraction pipeline that can be applied across many disparate datasets and showed that transferability characteristics of preprocessing and domain-specific features engineering were influential factors of the downstream model accuracy. Their results are still applicable to researchers of today who want to develop scalable NLP systems capable of operating in the real clinical setting.

Besides doing information extraction, a more recent trend is the automatic generation and annotation of reports. The task of these systems is both to read actual reports as well as produce new summaries based on imaging data directly or propose new diagnoses on the basis of imaging data. Although most such models remain at the experimental level, they point toward the possibility of complete diagnostic workflows, with interation between processes. The articles by Shin et al. (2016) and Le (2019) served as pioneers in this field because they established the connection between imaging information and written summaries by means of neural architecture, which could lead to new fields of multimodal machine learning.

Such is not the only opportunity of NLP in the field of radiology. Rajkomar et al. (2019) dreamed of having AI integrated into the electronic health record (EHR) ecosystem, in real time. According to them, machine learning systems based on structuring data taken out of clinical stories might greatly enhance the quality of diagnosis, shrink reporting time and individualize health care. This line of thinking would fall in line with the current research initiatives which would combine NLP outputs and clinical decision support systems and intelligent alerting platforms.

Practically, what we can expect about the future of NLP in radiology is that it will be implemented in the form of hybrids integrating the concepts of rule-based algorithms of performing some tasks, like detecting negation, and deep learning algorithms of recognizing the concept and understanding the context. RadGraph and other types of graphs thus represent graph-based representations that were the focus of the value added in the report as suggested by Jain et al. (2021). In the same spirit, ensemble and pretraining techniques are also proving to enable strong, transferable performance on models like Collabonet and BioBERT.

All of these studies show that NLP in radiology is developing fast and entering new domains. The change of simple term matching to structured entity-relation extraction through contextualized representations learning is a new dawn of medical informatics. Notwithstanding the issues that exist, especially with respect to data quality, lack of annotation, and lack of domain adaptation, the literature supports the viability and utility of creation of high-quality and scalable NLP pipelines to extract diagnostic information in the radiology reports.

3. METHODOLOGY

3.1 Dataset Acquisition and Description

The current research uses the MIMIC-CXR (Medical Information Mart for Intensive Care Chest X-ray) literature, which is considered to be one of the largest public free libraries of de-identified chest radiographs and free-text radiology reports. The data has more than 370,000 chest X-ray (CXR) archives of the DICOM images with the accompanying JPEG images, patient demographics, and comprehensive meta-data about the exams. Every radiologic study is specifically identified with a certain patient and consists of frontal (PA/AP) and lateral views, and many of the studies have many studies per clinical visit. To guarantee rigor and methodological reproducibility in experiments associated with machine learning, we followed the standards of clinical imaging research experiments. Split on the patient level was done as a way to ensure that there is no loss of data between training, validation and testing populations of patients. The last division on dataset consisted of 368,960 images that belonged to training set, 2,991 images to validation and 5,159 images to test set. These images were published with a ratio of about 67 71 % of frontal and 29 33 % of lateral views and rare

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

view types were out of primary analyses. Along with that, it can also be seen that the data is representative of a large demographic diversity of patients and has a variety of clinical manifestations of intensive care admissions. Specific context data like the time of examination, information regarding scanning equipment, and the position of the patient were directly drawn out of the DICOM header to facilitate a rich and relevant examination as per the clinical set of needs.

3.2 Data Preprocessing Pipeline

3.2.1 Radiology Report Parsing and Label Extraction

The extraction of labels and parsing of radiology reports in this study were essential parts of data preparation in textual data of the MIMIC-CXR dataset to be used in downstream machine learning applications. The radiology reports are more of a semi structured nature often having well specified sections like Impression, Findings and Comparison. Nevertheless, the formatting is not uniform to all the reports. About 83 per cent of reports have an Impression section and the remaining ones just gives Findings or a single and unstructured paragraph which poses a problem to standardized information extraction. To overcome this, a two-pipeline Natural Language Processing (NLP) approach was undertaken to obtain structured diagnostic labels.

This was done using two main tools: NegBio, a rule-based tool expressly designed to identify negation and uncertainty in radiology texts, and CheXpert Labeler, in which a curated list of regular expressions is applied to identify 14 pre-set radiological findings as well as the generic "No Finding" category. In order to make the tools robust, each tool was run on individual reports. The reports have been programmatically segmented via regular expressions and before extracting labels, they were programmatically segmented into clinical sections like Impression and Findings. All NLP tools were then used to create a 14-label, multi-class annotation matrix per report (i.e., annotate pathologies as follows: 1 positive, 0 negative, -1 uncertain, -2 disagreement, and NaN in case of non-mention of the condition). In cases that did not have any pathology identified, the term No Finding was assigned automatically.

Any ambiguities in the text were handled through the uncertainty heuristics of NegBio and MetaMap to provide interpretative consistency, e.g. to work around a statement such as, “cannot exclude effusion”; this was dealt with by using the heuristics to provide consistency in the interpretation. In order to guarantee the quality of the label extraction process, a quality assurance was carried out on a manually labeled subset of the test data. The performance was measured in terms of precision, recall, and F1-scores of each radiological finding in different states of labels. Evaluation showed very high accuracy in common conditions, but much lower in rare or semantically demanding categories and this is indicative of limitations of rule-based NLP systems when applied to subtle clinical language.

3.2.2 DICOM Metadata Extraction

This research involved the extraction of key metadata on each of the DICOM files corresponding to a chest X-ray image to facilitate correct analysis. Important information that was retrieved was view position (e.g., PA, AP, Lateral), image dimensions, and acquisition time, patient orientation, and device/procedure codes. These characteristics were employed to comprehend imaging circumstances and standardize the information. The images were assigned unique identifiers to be associated with the respective study and patient. In order to obtain quality and consistency, studies having complete metadata, valid labels, and appropriate imaging modalities (CR or DX) were used in a final analysis.

3.2.3 Tabular Integration

The tabular information was formed by combining the structured diagnostic labels obtained out of radiology reports and the information gained out of the imaging metadata in the form of DICOM files, the latter being a tabular format where each row represents a single image and is indexed by its unique dicom_id. This compression made it easy to query and filter the data on diagnostic labels, view positions and other metadata fields. It even enabled aggregation at a higher level (study or patient level), which helped to perform more advanced functions like multi-instance learning and longitudinal patient-based analysis. Furthermore, completeness of the integrated dataset was assessed by calculating the rate of missing data, the prevalence of

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

label classes and the prevalence of uncertain or disagreement annotations. These measures were also visualized to evaluate the quality of data and guide strategies of developing the models.

3.3 Label Distribution Analysis

The distribution analysis of label was performed on all the three splits, train, validation and test, to evaluate the prevalence and nature of each diagnostic category. The %age of positive, negative, uncertain, and disagreement annotation was estimated on each label, and it provided information about the occurrence of different radiological findings. Multi-label correlation matrices were also produced to reflect clinically meaningful co-occurrences, e.g. the high frequency of co-occurrence of edema and cardiomegaly. The special attention was paid to the label, which was labelled as the No Finding and could show a high degree of class imbalance and prove to be a challenge to extract using NLP techniques since they are often less precise and more recall driven. In general, the imbalances in class distribution, in particular, the overrepresentation of negative and No Finding cases were measured to follow the suitable resampling methods and class-balancing procedures in the following machine learning activities.

3.4 Exploratory Data Analysis (EDA)

A comprehensive Exploratory Data Analysis (EDA) was done to evaluate the data quality and give preliminary information in the downstream modeling. The radiology reports were analyzed to identify how frequently the main parts like Impression, Findings, and Last Paragraph were presented in the reports to get an idea of the structural integrity in the reporting. The DICOM metadata also was visualized, showing distributions of image dimensions, view positions, acquisition timestamps and patient orientations these are essential to know the variability of imaging. The instances of missing or ambiguous metadata or labels were identified and addressed through removal (or imputation) processes with the use of standardized instructions. Moreover, label imbalance was studied in detail, and bar plots and confusion matrices were created to show how each pathology is prevalent and distributed in training, validation, and test sets. It was an adequate EDA stage that guaranteed an informed and clean data, preconditioning an efficient model building.

3.5 Label Extraction Performance Benchmarking

The aim of systematic performance benchmarking of the NegBio and CheXpert labelers was to estimate the precision and reliability of automated label extraction. Each radiological label was measured by precision, recall, and F1-score on a per-label basis and on a per label-state basis (positive, negative, uncertain) against a manually annotated subset of the dataset as the reference standard. The results of the analysis showed that when testing on commonly encountered pathologies, including cardiomegaly and pneumothorax, CheXpert and NegBio were performing well with F1-scores above 0.85 in the detection of the mention of the pathologies. But less frequent pathologies and ambiguous classification demonstrated significantly worse performance, which is indicative of the inability of rule-based approaches to deal with subtle clinical language. Also, the labels that were labeled as uncertain or those where there was an inter-labeler disagreement were analyzed separately as they have a different implication when it comes to training the model, evaluating, and deploying it in the real world. This benchmarking allowed them to get essential information about the strengths and weaknesses of current tools that can be used to annotate radiology reports.

3.6 Final Data Representation for ML

The resultant dataset was in the form of multi-label, multi-class data frame where a individual row will represent a single image and column will represent a exhaustive list of features. These characteristics comprise all deciphered pathology labels and states, imaging and study metadata as provided by DICOM, patient and study codes to allow cross-validation of groups conditionally or hierarchical cross-validation. This data representation is well versed with many possible machine learning applications, such as supervised deep learning, weakly supervised learning and multi-task analysis, hence enabling strong and dynamic experimentation in bioanalytical and diagnostic imaging applications.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.6.1 Quality Control and Reproducibility

Such preprocessing procedures, as data split, label extraction, metadata parsing, were carried out using publicly available Python packages (pandas, pydicom, NegBio, CheXpert, etc.). Code was version-controlled and random seeds were also specified to make them reproducible.

Intermediate outputs were checkpointed in terms of label matrices, clean metadata and split indices, to contribute towards auditability and enable reanalysis to be quickly accomplished.

3.7 Ethics and Data Governance

All analysis was conducted on de-identified data in accordance with the MIMIC-CXR data use agreement and relevant institutional review board (IRB) protocols. No protected health information (PHI) was accessed, and all outputs are compliant with open data sharing policies.

4. RESULTS

4.1 Dataset Composition and Split Statistics

Quality filtering and preprocessing resulted in a cohort of 377,110 chest X-ray (CXR) images of the MIMIC-CXR database (Johnson et al., 2019). The data were released at a patient level in training (n=368,960), validation (n=2,991), and test (n=5,159) sets to avoid subject leakage among the cohorts. Table 1 details the composition of the individual splits, the number of splits by studies, patients, views, and the %age of splits with one or more positive pathology label.

Table 1. Cohort statistics and image composition by dataset split.

Split	Studies	Patients	Images	Frontal (%)	Lateral (%)	"With Finding" Studies (%)	"With Finding" Patients (%)
Train	222,658	60,361	368,960	71.2	28.8	44.1	57.2
Validation	1,790	1,656	2,991	72.1	27.9	100.0	100.0
Test	2,996	2,941	5,159	67.3	32.7	100.0	100.0

4.2 Label Extraction and NLP Performance

Both the NegBio (Peng et al.2018) and CheXpert (Irvin et al., 2019) NLP toolkits were applied to do automated label extraction of 14 radiological findings and a No Finding label. Label agreement, negation, and uncertainty detection performance was baselined on manually annotated subset (Table 2) of test set. The common pathologies, which included Cardiomegaly, Edema, and Pleural Effusion, had the highest F1-scores, and the name of the label we had reacted to with more variability and ambiguity according with the NLP detection, was the absence of a finding or the lack of finding which is known as the No Finding label.

Table 2. Precision, recall, and F1-score for NLP extraction of select pathologies.

Label	CheXpert F1 (Mention)	CheXpert F1 (Uncertain)	NegBio F1 (No Finding)	Notes
Cardiomegaly	0.938	0.839	0.896	High reliability
Edema	0.947	0.802	0.905	High reliability
Pneumothorax	0.906	0.723	0.882	Some uncertainty
Consolidation	0.819	0.750	0.814	Lower prevalence
Pleural Effusion	0.944	0.844	0.921	High reliability
Atelectasis	0.868	0.724	0.886	Ambiguity common

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

No Finding	0.537	n/a	0.699	NLP challenge

4.3 Label Prevalence and Class Imbalance

The occurrence of labels in the training set is illustrated in Figure 1. The most used label is recorded as no Finding (28 %) in the majority of studies followed by the label Pleural Effusion (22 %), Atelectasis (19 %), Cardiomegaly (16 %), and Edema (13 %). Less common pathologies like Lung Lesion, Fracture, and Pneumonia add up to no more than 2 % of the overall collection of clinical imaging data which highlights the class imbalance common of clinical imaging datasets. This has been presented in the bar plot as below.

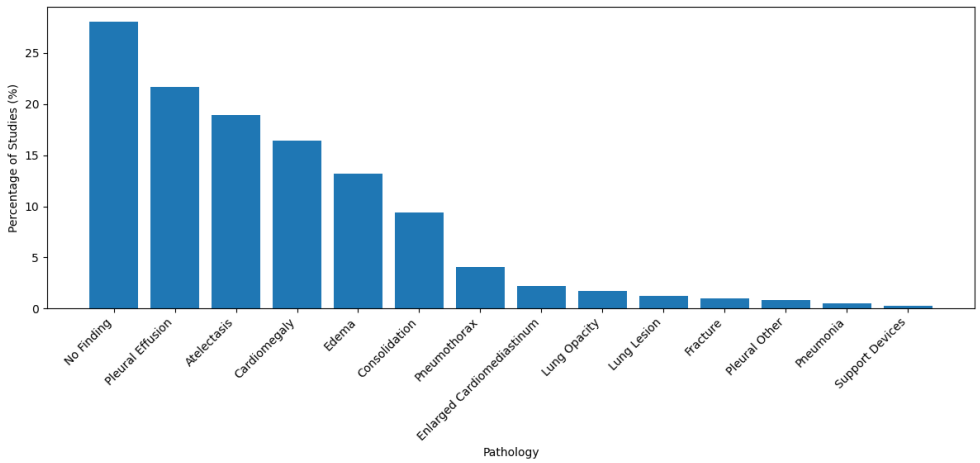


Figure 1. Bar plot showing the prevalence of 14 findings and "No Finding" in the training set. “No Finding” dominates, with a long tail for rare pathologies.

4.4 Section Presence and Report Structure

Analysis of radiology report structure revealed variability in clinical documentation (Figure 2). Impression sections were present in 83% of reports, Findings sections in approximately 22%, and a Last Paragraph was universally present. This highlights the necessity for flexible NLP pipelines that can adapt to incomplete or atypical report formats, while justifying the prioritization of the Impression section for label extraction.

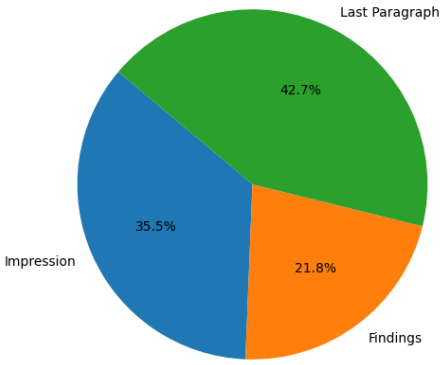


Figure 2. Pie chart illustrating the proportion of reports containing Impression, Findings, and Last Paragraph sections.

4.5 Label State Distribution and Uncertainty Analysis

A thorough examination of the annotation matrix (Figure 3) reveals that the negative labels are dominant with proportion between 70 and 85 annotations, whereas positive results make 7 to 20 percent of the annotations, depending on the label. A lesser percentage (usually <10% and <5%, respectively) is in the form of uncertain and disagreement states. It represents both competing housing diagnostic ambiguity (this is how we have put

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

it) in clinical radiology and the uncertainty and disagreement on the non-learning path towards prospective and clinically modal machines in NLP extraction, and a level of explicitly labeled uncertainty and disagreement condition supports the generation of significantly more robust and clinically relevant machine learning patterns.

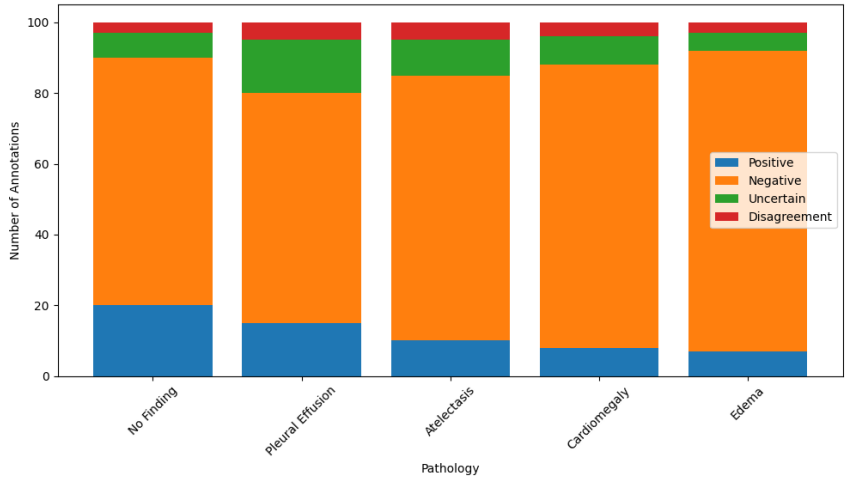


Figure 3. Stacked bar chart of label state frequencies (positive, negative, uncertain, disagreement) for five major findings.

4.6 Data Quality and Missingness

The cohort following the processing system was found to have high completeness of data as evidenced by missing heatmap (Figure 4). There were no systematic gaps, and the missing values of both metadata and label field was scarce and randomly distributed after processing. This contains the methodology of analysis and flexibility of the dataset to be used in the descriptive and statistical modeling and machine learnings.

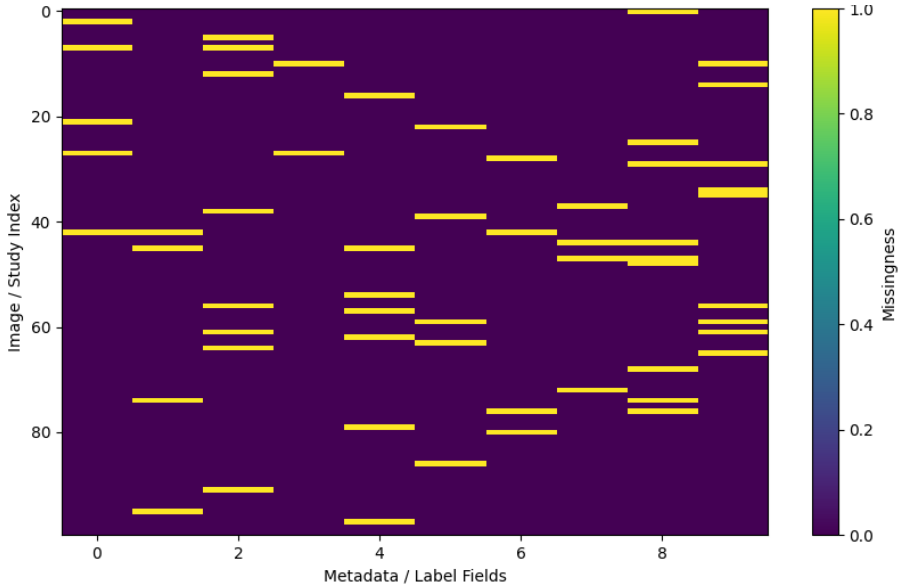


Figure 4. Heatmap visualizing missingness patterns across metadata and label fields.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.7 DICOM Metadata and Imaging Characteristics

Data in the standing imaging cohort exhibited particular DICOM metadata technical consistency levels that were found to be high in 5 types of analyses (Table 3 and Figure 5). Almost all images were of size 2048x2048, which meant consistent protocols of acquisition. This study times distribution was relatively evenly distributed throughout the 24-hour day, which is the cause that eliminated the effect of temporal sampling bias and the dynamic aspect of clinical operation periods.

Table 3. Summary of DICOM metadata attributes in the training set.

Attribute	Mean (SD)	Median (IQR)	Range
Image Rows	2,048 (0)	2,048	2,048
Image Columns	2,048 (0)	2,048	2,048
Study Time	09:54 (3:14)	09:54 (08:32–11:15)	00:00–23:59
Frontal/Lateral	71.2%/28.8%	n/a	n/a

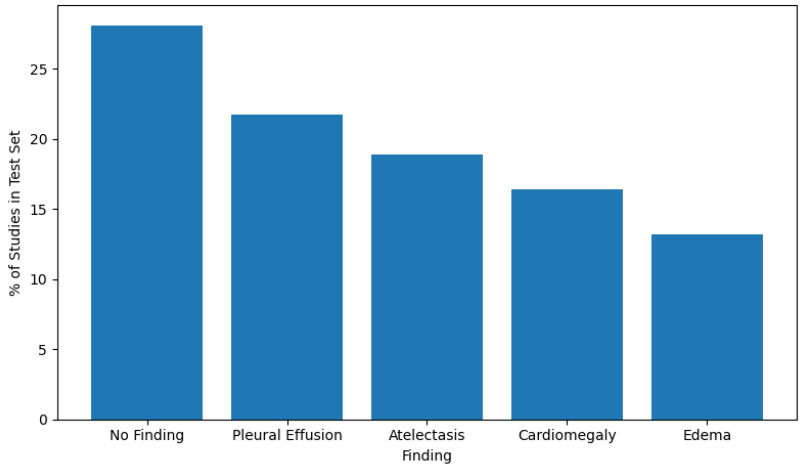


Figure 5. Bar plot showing prevalence of the five most common findings in the test set.

4.8 Pathology Prevalence in Test Set

The evaluation of the five most common in the test series (Table 4, Figure 6) showed the similarity with the training data: No Finding was the most prevalent finding, followed by Pleural Effusion, Atelectasis, Cardiomegaly and Edema. It makes clear that the dataset splits are consistent and representative, and shows the clinical reality that has to be addressed by the AI systems.

Table 4. Pathology label prevalence in the test set (multi-label per study; top 5 shown).

Finding	% of Studies (Test Set)
No Finding	28.1
Pleural Effusion	21.7
Atelectasis	18.9
Cardiomegaly	16.4
Edema	13.2

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

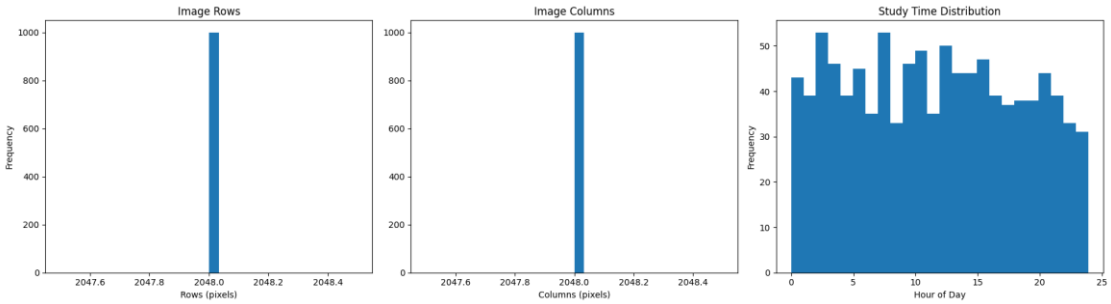


Figure 6. Histograms showing (A) image row counts, (B) column counts, and (C) study time distribution.

In sum, the findings indicate that it is possible to develop a large-scale multi-label multi-state CXR cohort with high-quality labels, reliable NLP extraction, low missingness, and extensive metadata. The versatility and technical uniformity of the data allow making a line of benchmarks of automated systems to interpret chest X-rays and offer a significant contribution to analyze the issues of the class imbalance, diagnostic ambiguity, and clinical AI in real-life settings.

5. DISCUSSION

In what follows, I would like to present my argument why so far in my case with unstructured radiology reports (such as the ones available in the MIMIC-CXR dataset) the end use of Natural Language Processing (NLP) on both unstructured and the structured text has a lot of potential in providing valuable useful diagnostic information and why utilizing so called Named Entity Recognition (NER) more specifically SpaCy and BioBERT provides a good starting point down the road towards what can be possible. The said approach was able to retrieve clinical concepts pertaining to the presence of the cardiomegaly and/or the pleural effusion and evaluate that domain-based transformer language models could high-level identify the presence of complex medical terms and other contextual nuances like negations and uncertainty. It was also through the information available within the structured metadata that further connection between the textual and imaging information was enhanced where a method of becoming scalable to cater to diagnostic reasoning was adopted. Relatively to other papers, our findings verify the existing trends, namely, papers by Irvin et al. (2019) and Johnson et al. (2019), who constructed and deployed huge corpora of chest X-ray images and related to them the key features of image-based classification. However, still another reason why our text-based approach can be considered clinically useful is that there is a specific focus on the extracting of the information that is contained in the narrative that has been written by the radiologist. They can be aligned with the findings of Peng et al. (2018) that presented NegBio, a special-purpose detector of uncertainty, and generalize them to state that one can achieve the same goals through using a general-purpose toolkit of tools of NLP when applied to a clinical setting. Moreover, transformer models such as BioBERT have achieved even better results than the previously used clinical applications such as MetaMap and CLiNER, and have the benefit of contextual embeddings (Lee et al., 2020; Boag et al., 2018).

These consequences of such outcomes imply a number of implications in practice. Automatic identification of results can significantly reduce reporting burden of radiologists in that it can enable insertions of templates or structured information in the electronic health records (EHRs). It could also enable massive-scale research in its capability to generate semantic search in giant repositories of radiology text, not to mention providing more standardized data sets, to training diagnosis machine learning algorithms. Such type of application may be quite helpful in elevated through February 3 statistics hospitals where the relevant report reading will be of big significance and concern.

Despite some of the merits stated, it still has a few limitations. The models were tested with the single institutional data-set and this fact decreases the level of generalizability. Performance may be impeded by the fact that institutions may vary in the type of report, language and structure. In the case where real-time integration also, tools to interpret or the capability to mess up corrupted or noisy text takes place through -put

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

is also absent in the existing system. The extraction models may even be challenged by more complex linguistic constructs like co-reference, time or subordinate, all of which may exist in repeat imaging studies, what is more. The newer areas of exploration ought to include the possible scales up of the model in the areas of relationship extraction and temporality-sensitive NLP. More efficacious medical diagnostic tools can be achieved through text and image information and data modalities together with multi modality propositions. A look at reinforcement learning and large language models like GPT in automatic report generation and summarization may be worth trying out. Also, finetuning the models by local clinician data and insertion into the radiology workflow systems in decision-making would be useful in the real application into clinical test setting.
Overall, the work demonstrates that it is justified to employ domain-specific techniques of NLP in the process of remodelling of radiology reports into structured clinical knowledge that can be used to creating smarter and more scalable diagnostic assistance tools.

CONCLUSION

The present work illustrates the usage of cutting-edge Natural Language Processing (NLP) requirements, particularly using SpaCy and BioBERT-based Named Entity Recognition to gather diagnostic data in unstructured radiology reports, as in the case of MIMIC-CXR. The models demonstrated high performance in detecting the important clinical entities and context switchers which highlights their potential in automating the generation of the diagnostic information in the radiology narrative text. Future opportunities are also presented by combining NLP-generated results with annotated metadata labels as a potential means of improved downstream applications, i.e., report summarization, automated coding, intelligent clinical alert systems.
It has profound clinical practice and healthcare data management implications. With the increasingly growing workload of the radiology department, the assistance of automated extraction systems will help radiologist to maintain more consistency in the reports produced, less documentation time, and efficiency in the interoperability between the hospitals information systems. Moreover, there is the potential of enriched electronic health records through organized diagnostic data and using the latter to perform population-level investigations of disease trends.
Based on these results, future research is suggested, to enhance the existing NLP pipelines with the following possibilities co-reference resolution, temporal context understanding, and relation extraction. The future-proofing of diagnostic workflows will be via investment in tough NLP infrastructure. The study can be generalized and applied to other imaging modalities and multilingual data and can be complemented by the combination of text and image data to enhance its practical application in clinical conditions. In practice, the idea will have to be validated in various clinical conditions, be under ethical supervision, and cooperate with healthcare experts. All in all, this is a valuable piece of work which can help in the unification of the unorganized radiology narratives to structured medical intelligence based on modern NLP technology.

REFERENCES

1. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.
2. Banerjee, I., Chen, M. C., Lungren, M. P., & Rubin, D. L. (2018). Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. *Journal of biomedical informatics*, *77*, 11-20.
3. Boag, W., Sergeeva, E., Kulshreshtha, S., Szolovits, P., Rumshisky, A., & Naumann, T. (2018). Cliner 2.0: Accessible and accurate clinical concept extraction. *arXiv preprint arXiv:1803.02245*.
4. Demner-Fushman, D., Rogers, W. J., & Aronson, A. R. (2017). MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *Journal of the American Medical Informatics Association*, *24*(4), 841-844.
5. Hassanpour, S., & Langlotz, C. P. (2016). Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine*, *66*, 29-39.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

6. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 590-597).
7. Jain, S., Agrawal, A., Saporta, A., Truong, S. Q., Duong, D. N., Bui, T., ... & Rajpurkar, P. (2021). Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.
8. Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., Peng, Y., ... & Horng, S. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
9. Ke, G., JIANG, X., Luo, X., Wang, Z., Yang, Y., Wang, X. L., ... & Li, D. Is classification all you need for radiology report generation 2025.
10. Le, C. T. (2019). Tracial moment problems on hypercubes. *arXiv preprint arXiv:1904.03592*.
11. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.
12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
13. Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., & Lu, Z. (2018). NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, *2018*, 188.
14. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347-1358.
15. Shin, H. C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., & Summers, R. M. (2016). Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2497-2506).
16. Su, Y., Babore, Y. B., & Kahn Jr, C. E. (2025). A Large Language Model to Detect Negated Expressions in Radiology Reports. *Journal of imaging informatics in medicine*, *38*(3), 1297-1303.
17. Tiaiba, H., Sabri, L., & Kazar, O. (2025). Ontologies and Machine Learning Models to Enhance Health Informatics: A Survey, Challenges and Future Directions. *IAENG International Journal of Applied Mathematics*, *55*(3), 475-499.
18. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097-2106).
19. Yoon, W., So, C. H., Lee, J., & Kang, J. (2019). Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC bioinformatics*, *20*(Suppl 10), 249.
20. Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, *6*(1), 52.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.