Received: 23 August 2022 | Accepted: 23 February 2023

DOI: 10.1002/hcs2.40

REVIEW

HEALTH CARE SCIENCE

Automated labelling of radiology reports using natural language processing: Comparison of traditional and newer methods

Seo Yi Chng[1] | **Paul J. W. Tern**[2] | **Matthew R. X. Kan**[3] | **Lionel T. E. Cheng**[4]

[1]Department of Paediatrics, National University of Singapore, Singapore, Singapore

[2]Department of Cardiology, National Heart Centre, Singapore, Singapore

[3]NUS High School of Mathematics and Science, Singapore, Singapore

[4]Department of Diagnostic Radiology, Singapore General Hospital, Singapore, Singapore

Correspondence
Seo Yi. Chng, Department of Paediatrics, National University of Singapore, Singapore, 5 Lower Kent Ridge Rd, Singapore 119074.
Email: drchngsy@gmail.com

Funding information
None

Abstract

Automated labelling of radiology reports using natural language processing allows for the labelling of ground truth for large datasets of radiological studies that are required for training of computer vision models. This paper explains the necessary data preprocessing steps, reviews the main methods for automated labelling and compares their performance. There are four main methods of automated labelling, namely: (1) rules-based text-matching algorithms, (2) conventional machine learning models, (3) neural network models and (4) Bidirectional Encoder Representations from Transformers (BERT) models. Rules-based labellers perform a brute force search against manually curated keywords and are able to achieve high F1 scores. However, they require proper handling of negative words. Machine learning models require preprocessing that involves tokenization and vectorization of text into numerical vectors. Multilabel classification approaches are required in labelling radiology reports and conventional models can achieve good performance if they have large enough training sets. Deep learning models make use of connected neural networks, often a long short-term memory network, and are similarly able to achieve good performance if trained on a large data set. BERT is a transformer-based model that utilizes attention. Pretrained BERT models only require fine-tuning with small data sets. In particular, domain-specific BERT models can achieve superior performance compared with the other methods for automated labelling.

KEYWORDS
automated labelling, machine learning, natural language processing, neural network, radiology

Abbreviations: BERT, Bidirectional Encoder Representations from Transformers; LSTM, long short-term memory; NLP, natural language processing; RNN, recurrent neural network; TF-IDF, term frequency-inverse document frequency.

1 | INTRODUCTION

Computer vision, with object detection or image segmentation algorithms, is gaining prominence in its ability to augment radiologists in interpretation of scans. Large, labelled data sets of many thousands of images are needed for the development of object detection or segmentation models using machine learning or deep learning. However, it is laborious for expert radiologists to manually label ground truth in such a large collection of images [1]. Employing natural language processing (NLP) to automatically label radiology reports presents a solution to this problem [2].

NLP is a branch of Artificial Intelligence (AI) in which computers are able to intelligently understand, interpret and generate text. NLP systems need to be able to overcome ambiguity in natural language to reliably extract information and interpret phrases in a context-dependent manner [3]. Specifically, biomedical texts present an additional challenge with its own set of terms, acronyms and shorthand that differs from the structure of typical language [4].

Automated labelling requires parsing the text of a radiology report to extract information about the presence or absence of prespecified pathologies. Within this remit, negation handling is especially important [5]. Various NLP techniques have developed over the years, with increasing ability to take context into account for greater accuracy of information extraction. Accordingly, newer models of NLP are increasingly complex, building on their predecessors. This paper aims to review the major NLP techniques involved in automated labelling, starting from (1) a simple rules-based text-matching method, (2) the basic conventional machine learning model, (3) a neural network model that affords greater memory for context and (4) the Bidirectional Encoder Representations from Transformers (BERT) model, which incorporates the concept of attention and calculates a context vector to pass along information about how much significance should be assigned to each input word.

Within the medical context, beyond automated reports, NLP and its applications to medicine remains an important field of research. NLP models can be used to extract, aggregate and condense information from electronic health records both for research purposes and for routine clinical care [6]. It has also been employed in summarizing key insights from the corpus of biomedical literature. More recently, Large Language Models have been developed, which are large pretrained AI systems that can be fine-tuned and repurposed to specific tasks, such as mimicking patient–doctor interactions [7] or answering clinical questions [8].

2 | DATA SETS

Although there are a number of publicly available radiology image data sets available, there are only a few data sets that contain full radiology reports. MIMIC-CXR Database v2.0.0 [9, 10] and the Indiana University Chest X-Ray Collection [11] are two notable data sets that contain full radiology reports. The MIMIC-CXR Database contains more than 200,000 radiology reports, whereas the IU X-ray data set contains almost 4000 radiology reports from two large hospital systems. There is a paucity of publicly available radiology reports for other imaging modalities (ultrasound, computed tomography, magnetic resonance imaging and so on) and other anatomical locations other than the chest.

Human-labelled reports are important as ground truth for determining the performance of the various methods for automated labelling of radiology reports [12]. However, in radiology, as with all other medical fields, subject matter experts may be both difficult to find and expensive to engage [13]. To address this problem, some researchers have attempted to use weak supervision to label the data in place of humans [14]. In weak supervision, a set of labelling functions in the form of rules, dictionaries, ontologies and pretrained machine learning or deep learning models may be used. A statistical model takes the labels from the functions as input and outputs the probabilistic labels to be used as ground truth. By adding more functions, the labelling performance would improve, but it may not be able to achieve the same performance as labels generated by expert radiologists. Hybrid models that use weak annotations together with few strongly annotated labels have been implemented as an improved means of labelling the data, rather than using weak annotations alone [15].

3 | DATA PREPROCESSING

The text in radiology reports requires preprocessing to ensure that the data is in a suitable format for feeding into conventional machine learning models for multi-label classification. The report is first segmented into individual sentences, and then tokenized into individual words. All text is also converted into lowercase as this does not affect their meaning. Stop words, which connect keywords together to form coherent and grammatically correct sentences, but have no intrinsic relevance and meaning, are removed. However, negative words such as ‘no’ and ‘not’ are retained.

Negation handling is an important aspect [5] in the automated annotation of radiology reports. There are

several approaches, which can be used to handle negation in medical texts for machine learning classification tasks. A simple way would be to concatenate the prefix 'no_' with every word in a sentence in which a negative word is detected. Alternatively, specialized dependency parsers such as the Negbio package [16], which was trained on Biomedical text, can be utilized to account for negation in radiology reports.

As most machine learning algorithms are unable to process text, the text data has to be vectorized. Vectorization (also known as word embeddings) is the process of converting text data into numerical vectors [17]. The data are converted into a matrix of term frequency-inverse document frequency (TF-IDF) features, whereby the top features are incorporated and ordered by term frequency. TF-IDF is a measurement of the importance of a word, with the idea being that a term with more significance in a particular report will occur more frequently in that context as compared with its relative frequency across the entire corpus of reports. N-grams, which consider a sequence of n words in the text, can be used as potential features. N-grams are preferred to other vectorization techniques such as Bag of Words.

GloVe [18] word embeddings that were developed using Radiopedia as a general radiology corpus [19] can also be used to convert text into vectors. GloVe takes into account global word–word co-occurrences in the entire corpus.

Special data processing is required to apply the pre-trained BERT [20]. The BERT tokenizer provided by the library must be used, as the BERT model has a specific and fixed vocabulary and the BERT tokenizer is able to handle out-of-vocabulary words. The BERT tokenizer splits the radiology text into tokens, adds the required tokens of [CLS] and [SEP] to the start and end of each sentence, and then converts the tokens into indexes of the tokenizer vocabulary. The BERT tokenizer then pads or truncates all sentences to a single constant length, and lastly creates an attention mask.

4 | EVALUATION METRICS

In automated annotations of radiology reports, the F1 score is frequently used to evaluate the success of the classifier [21, 22, 23]. The F1 score is the harmonic mean of precision and recall, and measures the classifier's performance for both positive and negative cases. Precision is the fraction of true positives divided by the total number of test positives. Recall is the fraction of true positives divided by the total number of disease positives. The classification report is also frequently generated to display the precision, recall and F1 score for each class label. Similarly, the area under the precision recall curve score can also be used as an evaluation metric to compare the performance of various models.

The accuracy score should preferably not be used as an evaluation metric [24] as the datasets used are typically imbalanced (with the minority class occurring <40% of the time). Most disease labels will be negative in a radiology report. The accuracy score can be misleadingly high for certain classes if a model predicts negative 100% of the time [24]. Similarly, the area under the receiver operating characteristic curve score can be misleadingly high in imbalanced data sets if the model predicts the majority class 100% of the time.

5 | RULES-BASED TEXT-MATCHING ALGORITHM

Labelling of radiology reports begins with the identification of related keywords that map onto a particular label. Table 1 shows a sample list of relevant keywords for common class labels in chest X-ray reports. The most basic form of automated labelling involves a text-matching algorithm, which searches the report text for predefined keywords and assigns corresponding labels based on certain rules.

TABLE 1 Common class labels in chest X-ray reports with a sample list of relevant keywords.

Label	Related keywords
Air-space opacity	Consolidation, ill-defined increased parenchymal opacity, obscured vascular markings, pneumonia
Nodular opacity	Nodule, nodularity, cannon-ball lesion
Interstitial opacity	Reticular opacity, reticulation, reticulo-nodular opacity, honeycombing
Cardiomegaly	Enlarged cardiac silhouette, enlarged heart, increased cardiothoracic ratio
Pleural effusion	Blunted costophrenic angles, fluid in pleural space, pleural fluid
Pneumothorax	Gas in pleural space

Medical reports (which include clinical summaries, radiology reports and pathology reports) often contain negative words [5]. These negative words are used to communicate that important diagnoses have been excluded. For example, a sample radiology report may contain the following sentences: 'The heart is not enlarged. There is no evidence of oedema'. A rules-based text-matching algorithm that simply identifies the keyword 'oedema' in the labelling process while neglecting the negative modifier of 'not' or 'no' would lead to misclassification of the report.

It is crucial that rules-based text-matching algorithms identify related negative modifiers within the same phrase or sentence [5]. Figure 1 illustrates the typical logic flow in such algorithms.

Expert systems were one of the first AI tools to be developed for the automated annotation of radiology reports [21]. An expert system comprises two subsystems, namely the knowledge base and the inference engine. The knowledge base contains facts and rules, created with the specialized knowledge from human domain experts. In the field of Medicine, the inference engine applies the rules to known facts, and the output is used for medical diagnosis or to guide treatment decisions. As early as 1989, a knowledge-based data acquisition tool (Special Purpose Radiology Understanding System) was successfully implemented [21]. The system used semantic information from a diagnostic knowledge base to drive the understanding of chest radiology reports.

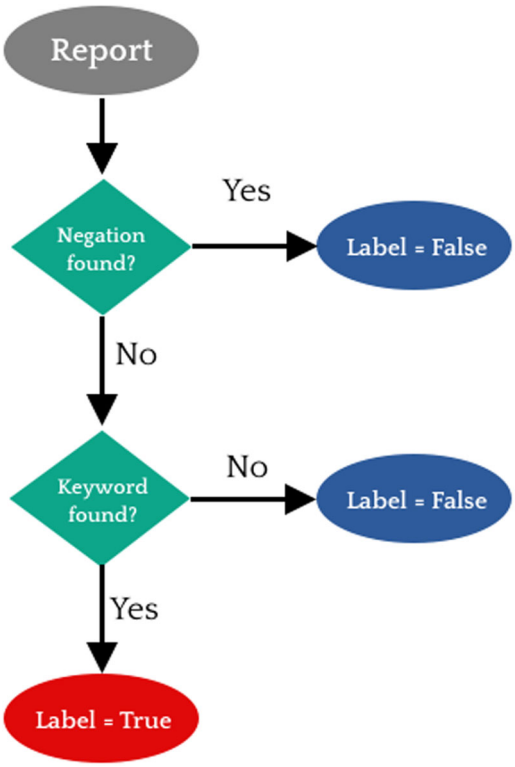

FIGURE 1 Example flow diagram of rules-based text-matching algorithms.

Rule-based labellers that brute-force search against manually curated keywords are technically easy to implement and do not require heavy computational power or memory. Furthermore, such classifiers can achieve unexpectedly high performance [22]. For example, the CheXpert rules-based classifier achieved state-of-the-art performance with the highest F1 scores of 0.743 (95% confidence interval [95% CI]: 0.719–0.764) among all the algorithms, including that of machine learning and deep learning models [22], with the exception of CheXbert [23].

6 | CONVENTIONAL MACHINE LEARNING MODEL

Although outpatient radiological investigations typically have normal findings, a single inpatient image may contain multiple lesions. This necessitates techniques that allow for multilabel classification. Possible multilabel classification approaches [25] include Binary Relevance, Classification Chain, Label Powerset and Multilabel K-Nearest Neighbour. Binary relevance is less suitable, as it ignores correlations between class labels, which occur frequently in radiology reports. For example, the labels 'cardiomegaly' and 'oedema' are correlated, as both findings occur together in congestive cardiac failure. Label Powerset provides a more optimal approach by transforming a multilabel classification problem into a multiclass problem, with one multiclass classifier trained on all unique label combinations in the training data. Label Powerset maps each combination to a unique combination identification number and performs multiclass classification using the combination identification numbers as classes.

Conventional machine learning models that can be employed as classifiers in automated labelling of radiology reports include Naive Bayes, Logistic Regression and Random Forest ensemble. Naive Bayes algorithm is a classic algorithm used in NLP [26]. It is a probabilistic algorithm based on applying Bayes theorem with the 'naive' assumption of conditional independence between every pair of features. The multinomial Naive Bayes algorithm is widely used for assigning documents to classes by calculating the probability of each label for a given text and outputs the label with the highest probability. Naive Bayes classifiers work better than expected, considering their naive design and simplified assumptions of independence. However, the simplicity of the Naive Bayes algorithm results in it having poorer

performance compared with ensemble methods such as Random Forest classifiers [27]. In addition, radiology class labels are frequently correlated and not independent, thus making the Naive Bayes classifier a less ideal choice.

Random Forest [28] is a versatile model and the data does not need to be rescaled or transformed. It is fast to train as the model works with only a subset of features. The Random Forest algorithm also balances errors in imbalanced data sets, which is especially helpful in labelling radiology reports as some labels occur infrequently. In a study on the automated annotation of 1295 magnetic resonance imaging reports of the knee, the F1 scores for the Naive Bayes and Random Forest classifiers were 73.8% and 82.2%, respectively [29].

Machine learning models are able to achieve high F1 scores if they have access to a large training set of reports that have been manually labelled by expert radiologists, with at least 1000 positive samples per class. However, if the data is insufficient, particularly in the case of uncommon labels, training will be compromised and the predictive model may be inaccurate [30].

7 | DEEP LEARNING (NEURAL NETWORK) MODEL

Recurrent neural networks (RNNs) [31] are ideal for sequential data [32] such as text and time series data, where the sequence is more important than the individual items. In NLP, the next word is commonly dependent on the previous words and there is a need to remember the previous words. Hence, as opposed to feed-forward neural networks (where the inputs and outputs are independent), the output from the previous step in RNNs is used as input to the current step. RNNs have a hidden state to capture information about a sentence. RNNs also have a memory that retains information about the calculations made so far for a period of time. A shortcoming of RNNs is that they are unable to handle long-range dependencies due to vanishing gradients [31].

Deep learning models for text classification typically use a long short-term memory (LSTM) network. LSTM [31] is a type of RNN that performs better than a conventional RNN in text classification tasks, as an LSTM network has superior memory ability and can handle long-term dependencies. LSTMs can overcome the vanishing gradient problem seen in RNNs [31].

LSTMs consist of three parts, the Forget gate [31], the Input gate and the Output gate. The Forget gate determines if the information from a previous period is relevant and to be remembered, or irrelevant and to be forgotten. The Input gate learns new information from the input. The Output gate passes the updated information to the next timestamp. Similar to RNNs, LSTMs have hidden states (short-term memory) of the previous timestamp and the current timestamp. LSTMs also have cell states (long-term memory) of the previous and current timestamps.

A possible deep learning model for this task of automated annotation of radiology reports would have 1 input layer, 1 embedding layer, 1 LSTM layer with 128 neurons and 1 output layer. The number of neurons in the output layer should correspond to the number of labels required in the output. Figure 2 shows an example of such a deep learning model.

In a study of automated annotation of 1295 magnetic resonance reports of the knee, a neural network model was able to achieve an overall superior F1 score of 0.867, compared with conventional machine learning models such as logistic regression (F1 0.846), random forest (F1 0.822) and naive Bayes (F1 0.738) [29]. However, as neural networks require large amounts of training data, the neural network model performed notably worse than the conventional machine learning models on the underrepresented classes with few training instances.

8 | BERT MODEL

The Transformer [33] is a novel NLP architecture that is able to undertake sequence-to-sequence tasks while accounting for long range dependencies. The Transformer

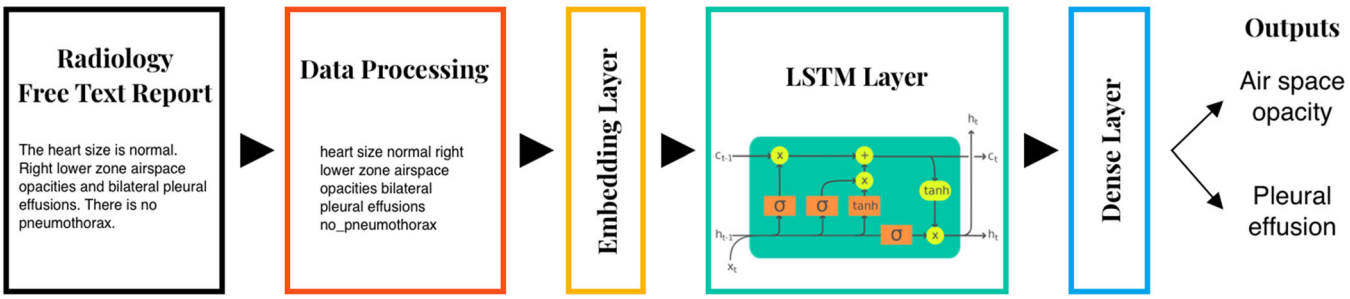

FIGURE 2 Example configuration of deep learning model. LSTM, long short-term memory.

utilizes self-attention to convert input sequences into output sequences without the need for RNNs or convolutional neural networks. Self-attention (also known as intra-attention) relates different areas of a sequence to compute a representation of the sequence. The Transformer performs multihead attention, where self-attention is computed both in parallel and independently many times. Although the Transformer is able to learn longer-term dependencies and is a marked improvement over RNN-based sequence-to-sequence models, the Transformer is limited as its attention can only handle fixed-length text strings. The input text has to be split into segments or chunks and such chunking results in context fragmentation. The context will be lost if the sentence is split in the middle.

The release of BERT [20] by Google AI heralded a new era in NLP. Although neural networks can be implemented for text classification, they need to be trained on a large number of manually labelled texts to achieve accurate results. In contrast, BERT has been pretrained on large quantities of unlabelled data by its developers and can subsequently be fine-tuned on a small number of manually labelled texts to achieve accurate results for a wide variety of tasks including text classification, language inference and question answering systems. The code for BERT is open-sourced and freely available for download and reuse, saving researchers computational time and resources from the need to repeat the training of a large data set.

BERT has substantially outperformed benchmarks in multiple NLP tasks. The superior performance of BERT can be attributed firstly to its use of transfer learning. As earlier explained, the base BERT model has been pretrained, allowing it to learn about text structure and language patterns. In addition, the BERT model is deeply bidirectional and is able to learn the meaning of the word in a sentence based on the words preceding and following the word, whereas other NLP models are unidirectional or shallowly bidirectional. Furthermore, BERT models have more encoder layers, larger feed-forward networks and more attention heads when compared to the original Transformer as a point of reference [20]. With BERT, each layer applies self-attention and the results are passed through a feed-forward network to the next encoder.

To utilize the pretrained BERT, the text has to be split into tokens, which are then mapped to indexes of the tokenizer vocabulary. All the sentences have to be padded or truncated to a single constant length and an attention mask must be created. During the training process, a BertClassifier class is created to extract the last hidden state of the classification [CLS] token and input it into a single layer feed-forward neural network to compute logits. Prediction is similar, with a forward pass to compute logits and calculate probabilities. A threshold of 0.5 can be used, such that sentences with a predicted probability >50% will be labelled as positive.

BERT classifiers have been shown to achieve superior performance [23, 34] compared with the other aforementioned techniques. CheXbert [23] is a biomedically pretrained BERT model that was initially trained on the outputs of a labeller and subsequently fine-tuned on manual annotations. It outperformed the previous best labeller, CheXpert, a rules-based labeller, in the labelling of chest X-ray reports. CheXbert was able to achieve state of the art performance with an F1 score of 0.798 (95% CI: 0.775–0.816) compared with the previous best performer CheXpert with an F1 score of 0.743 (95% CI: 0.719–0.764).

By adding a simple single-hidden-layer neural network classifier on top of BERT and fine-tuning BERT, superior performance can be achieved, even with small datasets with few positive labels. Additionally, as BERT was pretrained on vast amounts of text and has learned much language structure, fine-tuning BERT on radiology reports thus only requires a small single-digit number of epochs to achieve state-of-the-art performance. Training, validation, and testing times for a pre-trained BERT model are very fast. DistilBERT [35] was able to complete training and validation in <4 min and was subsequently able to infer nearly 70,000 radiology reports at a speed of 0.005 s per case [36].

Newer transformer architectures, such as RoBERTa [37], can perform better than BERT in annotating radiology reports [36]. Domain-specific BERT, such as PubMedBERT [38], can also outperform the original BERT model with higher area under the receiver operating characteristic curve in automated labelling of radiology reports [36]. PubMedBERT was generated by extending the pre-training of a BERT base model over a corpus of PubMed abstracts and full PubMed Central articles and is able to generalise to a variety of biomedical NLP tasks [38]. RadBERT [39] is the first BERT-based language model adapted for radiology and was released in June 2022. As RadBERT was trained on millions of radiology reports and is tailored to radiology, it was able to demonstrate improved performance of radiology NLP tasks compared with baseline BERT models [39]. The RadBERT models are available for use with a data usage agreement and are expected to be widely adopted and accelerate the use of AI in radiology research.

BERT models can also be used to build knowledge graphs. Knowledge graphs, also known as semantic networks, represents semantics by describing entities and their relationships. Knowledge graphs make use of ontologies for logical inference to derive implicit knowledge. Entities and relations can be represented as

TABLE 2 Comparison of the four methods used for automated labelling of radiology reports.

Method	Amount of training data required	Training time	Hardware requirements	Ease of implementation	Interpretability	Performance
Rules-based (e.g., CheXpert [21])	None required	None required	CPU	Easy	Classifier is very easy to interpret	F1 0.743 (95% CI: 0.719–0.764)
Conventional machine learning (e.g., random forest ensemble on MR reports of the knee [23])	++	++	CPU	Somewhat harder	Most models are easy to interpret	F1 0.822
Neural network (e.g., neural network on MR reports of the knee [23])	+++	+++	GPU	Harder	Model is harder to interpret	F1 0.867
BERT (e.g., CheXbert [22])	+ (To fine-tune a pretrained model)	++ (To fine-tune a pretrained model)	GPU	Harder	Model is harder to interpret	F1 0.798 (95% CI: 0.775–0.816)

Abbreviations: BERT, Bidirectional Encoder Representations from Transformers; CI, confidence interval; CPU, central processing unit; GPU, graphics processing unit; MR, magnetic resonance.

knowledge graph embeddings and then used in machine learning. Using a BERT model, Zhang et al. [40] extracted entities and relationships from unstructured radiology reports and constructed a knowledge graph. The authors then used the knowledge graph to build a 25-label classification system to extract chest X-ray labels from radiology reports.

We have detailed above the four main techniques that can be employed in the automated labelling of radiology reports. Table 2 compares the four techniques, in terms of the amount of training data required, the training time, hardware requirements, ease of implementation of the models, interpretability and performance.

9 | CONCLUSION

Automated labelling of radiology reports using NLP provides an efficient method of generating labels that are needed for object detection and image segmentation in radiological scans. There are four main techniques used for this, namely (1) a rules-based text-matching method, (2) a conventional machine learning model, (3) a neural network model and (4) a BERT model. Of note, BERT is a novel NLP architecture that utilizes transfer learning and is able to achieve state-of-the-art performance with minimal training data.

AUTHOR CONTRIBUTIONS

Seo Yi Chng: Conceptualization (equal); writing—original draft (lead); writing—review & editing (equal). **Paul J. W. Tern**: Writing—review & editing (equal). **Matthew R. X. Kan**: Writing—original draft (equal); writing—review & editing (equal). **Lionel T. E. Cheng**: Writing—review & editing (equal).

ACKNOWLEDGEMENTS

The authors did not receive any external assistance in the drafting of this manuscript. The authors did not receive funding for this work.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

No data sets were generated or analysed during the current study.

ETHICS STATEMENT

Ethical approval was not required for the current study.

INFORMED CONSENT

Informed consent was not required in this study.

ORCID

Seo Yi Chng https://orcid.org/0000-0001-8216-4022
Paul J. W. Tern https://orcid.org/0000-0001-5382-9881
Matthew R. X. Kan https://orcid.org/0000-0002-6688-3454

REFERENCES

1. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. Radiology. 2020;295(1):4–15. https://doi.org/10.1148/radiol.2020192224
2. Mozayan A, Fabbri AR, Maneevese M, Tocino I, Chheang S. Practical guide to natural language processing for radiology. Radiographics. 2021;41(5):1446–53. https://doi.org/10.1148/rg.2021200113
3. Yadav A, Patel A, Shah M. A comprehensive review on resolving ambiguities in natural language processing. AI Open. 2021;2:85–92. https://doi.org/10.1016/j.aiopen.2021.05.001
4. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. J Biomed Inform. 2015;57:28–37. https://doi.org/10.1016/j.jbi.2015.07.010
5. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp. 2001:105–9.
6. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. IEEE/ACM Trans Comput Biol Bioinf. 2019;16(1):139–53. https://doi.org/10.1109/TCBB.2018.2849968
7. Bao Q, Ni L, Liu J. HHH: An online medical chatbot system based on knowledge graph and hierarchical bi-directional attention. In: *Proceedings of the Australasian Computer Science Week Multiconference*. Melbourne, VIC, Australia: ACM; pp. 1–10. 2020. Available from: https://doi.org/10.1145/3373017.3373049
8. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. arXiv. 2022;1. http://arxiv.org/abs/2212.13138
9. Johnson AEW, Pollard T, Mark R, Berkowitz S, Horng S. The MIMIC-CXR Database [Internet] [cited January 25, 2023]. 2019. Available from: https://physionet.org/content/mimic-cxr/
10. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019;6(1):317. https://doi.org/10.1038/s41597-019-0322-0
11. Demner-Fushman D, Antani S, Simpson M, Thoma GR. Design and development of a multimodal biomedical information retrieval system. J Comput Sci Eng. 2012;6(2):168–77. https://doi.org/10.5626/JCSE.2012.6.2.168
12. Plank B. The 'problem' of human label variation: on ground truth in data, modeling and evaluation. arXiv preprint. 2022. arXiv:2211.02570.
13. Mandivarapu JK, Camp B, Estrada R. Deep active learning via open-set recognition. Front Artif Intell. 2022;5:737363. https://doi.org/10.3389/frai.2022.737363
14. Banerjee I, Li K, Seneviratne M, Ferrari M, Seto T, Brooks JD, et al. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. JAMIA Open. 2019;2(1):150–9. https://doi.org/10.1093/jamiaopen/ooy057
15. Agnikula Kshatriya BS, Sagheb E, Wi CI, Yoon J, Seol HY, Juhn Y, et al. Identification of asthma control factor in clinical notes using a hybrid deep learning model. BMC Med Inform Decis Mak. 2021;21(Suppl 7):272. https://doi.org/10.1186/s12911-021-01633-4
16. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. arXiv. 2017;2. http://arxiv.org/abs/1712.05898
17. Asudani DS, Nagwani NK, Singh P. Impact of word embedding models on text analytics in deep learning environment: a review. Artif Intell Rev. 2023:1–81. Epub 2023 February 22. https://doi.org/10.1007/s10462-023-10419-1
18. Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics. pp. 1532–43. 2014. Available from: https://aclanthology.org/D14-1162
19. Chen TL, Emerling M, Chaudhari GR, Chillakuru YR, Seo Y, Vu TH, et al. Domain specific word embeddings for natural language processing in radiology. J Biomed Inf. 2021;113:103665. https://doi.org/10.1016/j.jbi.2020.103665
20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. 2019;2. http://arxiv.org/abs/1810.04805
21. Ranum DL. Knowledge-based understanding of radiology text. Comput Methods Programs Biomed. 1989;30(2–3):209–15. https://doi.org/10.1016/0169-2607(89)90073-4
22. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell. 2019;33(01):590–7. https://doi.org/10.1609/aaai.v33i01.3301590
23. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv. 2020;3. http://arxiv.org/abs/2004.09167
24. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022;12(1):5979. https://doi.org/10.1038/s41598-022-09954-8
25. Bogatinovski J, Todorovski L, Džeroski S, Kocev D. Comprehensive comparative study of multi-label classification methods. Expert Syst. Appl. 2022;203. https://doi.org/10.1016/j.eswa.2022.117215
26. Schneider KM. Techniques for improving the performance of naive bayes for text classification. In: Gelbukh A editors. Computational linguistics and intelligent text processing. CICLing 2005. Lecture notes in computer science, vol 3406. Berlin, Heidelberg: Springer; 2005. https://doi.org/10.1007/978-3-540-30586-6_76
27. Sadman N, Tasneem S, Haque A, Islam MM, Ahsan MM, Gupta KD. Can NLP techniques be utilized as a reliable tool

for medical science? Building a NLP framework to classify medical reports, 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2020, pp. 0159–0166. https://doi.org/10.1109/IEMCON51383.2020.9284834
28. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324
29. Krsnik I, Glavaš G, Krsnik M, Miletić D, Štajduhar I. Automatic annotation of narrative radiology reports. Diagnostics. 2020;10(4):196. https://doi.org/10.3390/diagnostics10040196
30. Kokol P, Kokol M, Zagoranski S. Machine learning on small size samples: a synthetic knowledge synthesis. Sci Prog. 2022;105(1) https://doi.org/10.1177/00368504211029777
31. Schmidt Robin M. Recurrent neural networks (rnns): a gentle introduction and overview. arXiv preprint. 2019. arXiv:1912.05911. https://doi.org/10.48550/arXiv.1912.05911
32. Ostmeyer J, Cowell L. Machine learning on sequential data using a recurrent weighted average. Neurocomputing. 2019;331:281–88. https://doi.org/10.1016/j.neucom.2018.11.066
33. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. 2017;5. http://arxiv.org/abs/1706.03762
34. Bressem KK, Adams LC, Gaudin RA, Tröltzsch D, Hamm B, Makowski MR, et al. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics. 2021;36(21):5255–61. https://doi.org/10.1093/bioinformatics/btaa668
35. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. 2020;4. http://arxiv.org/abs/1910.01108
36. Tejani AS, Ng YS, Xi Y, Fielding JR, Browning TG, Rayan JC. Performance of multiple pretrained BERT models to automate and accelerate data annotation for large datasets. Radiol Artif intell. 2022;4(4):220007. https://doi.org/10.1148/ryai.220007
37. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. 2019;1. http://arxiv.org/abs/1907.11692
38. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare. 2022;3(1):1–23. https://doi.org/10.1145/3458754
39. Yan A, McAuley J, Lu X, Du J, Chang EY, Gentili A, et al. RadBERT: adapting transformer-based language models to radiology. Radiol Artif Intell. 2022;4(4):e210258. https://doi.org/10.1148/ryai.210258
40. Zhang Y, Liu M, Hu S, Shen Y, Lan J, Jiang B, et al. Development and multicenter validation of chest X-ray radiography interpretations based on natural language processing. Commun Med. 2021;1:43. https://doi.org/10.1038/s43856-021-00043-x

How to cite this article: Chng SY, Tern PJW, Kan MRX, Cheng LTE. Automated labelling of radiology reports using natural language processing: comparison of traditional and newer methods. Health Care Sci. 2023;2:120–128. https://doi.org/10.1002/hcs2.40