Journal of Digital Imaging (2020) 33:131–136
https://doi.org/10.1007/s10278-019-00271-7

ORIGINAL PAPER

Automated Detection of Radiology Reports that Require Follow-up Imaging Using Natural Language Processing Feature Engineering and Machine Learning Classification

Robert Lou[1] · Darco Lalevic[2] · Charles Chambers[2] · Hanna M. Zafar[2] · Tessa S. Cook[2]

Published online: 3 September 2019

Abstract
While radiologists regularly issue follow-up recommendations, our preliminary research has shown that anywhere from 35 to 50% of patients who receive follow-up recommendations for findings of possible cancer on abdominopelvic imaging do not return for follow-up. As such, they remain at risk for adverse outcomes related to missed or delayed cancer diagnosis. In this study, we develop an algorithm to automatically detect free text radiology reports that have a follow-up recommendation using natural language processing (NLP) techniques and machine learning models. The data set used in this study consists of 6000 free text reports from the author's institution. NLP techniques are used to engineer 1500 features, which include the most informative unigrams, bigrams, and trigrams in the training corpus after performing tokenization and Porter stemming. On this data set, we train naive Bayes, decision tree, and maximum entropy models. The decision tree model, with an F1 score of 0.458 and accuracy of 0.862, outperforms both the naive Bayes (F1 score of 0.381) and maximum entropy (F1 score of 0.387) models. The models were analyzed to determine predictive features, with term frequency of *n*-grams such as "renal neoplasm" and "evalu with enhanc" being most predictive of a follow-up recommendation. Key to maximizing performance was feature engineering that extracts predictive information and appropriate selection of machine learning algorithms based on the feature set.

Keywords Artificial intelligence · Binary classification · Follow-up · Machine learning · Natural language processing · Structured reporting

✉ Robert Lou
robert.lou@pennmedicine.upenn.edu

Darco Lalevic
lalevic@pennmedicine.upenn.edu

Charles Chambers
charles.chambers@uphs.upenn.edu

Hanna M. Zafar
hanna.zafar@pennmedicine.upenn.edu

Tessa S. Cook
tessa.cook@pennmedicine.upenn.edu

[1] Perelman School of Medicine at the University of Pennsylvania, 801 S 24th St #3, Philadelphia, PA 19146, USA

[2] Hospital of the University of Pennsylvania, Philadelphia, PA, USA

Hypothesis

Natural language processing's ability to detect follow-up recommendations is dependent on feature engineering that produces features with high mutual information and subsequent selection of machine learning algorithms. Because radiology reports tend to generate large feature spaces, algorithms that are appropriate for large feature sets will offer improved performance.

Background

Many patients have incidental findings that are suspicious of malignancy or indeterminate, which warrant follow-up imaging. Unfortunately, a large portion of patients do not get follow-up imaging [1, 2]. In particular,

for abdominal imaging at our tertiary care center, we have found that up to half of patients with recommended follow-up imaging did not complete any follow-up [3]. This represents a large population of patients that are at risk for adverse outcomes due to missed or delayed cancer diagnoses. There are many reasons why patients do not get follow-up imaging. For example, referring physicians may not address follow-up imaging because it does not require immediate action or because it is not related to the medical concern for which the original imaging was indicated. A solution to this issue is explicitly specifying follow-up recommendations in the report so that such patients can be monitored for completion [3]. While structured reporting could explicitly indicate the need for follow-up, the fact is many radiology reports remain unstructured or loosely structured [4]. Automated identification of follow-up recommendations in radiology reports would allow for automated tracking of patients requiring follow-up and help to decrease the number of patients who experience adverse outcomes due to missed follow-up. In this study, we develop an algorithm to automatically detect free text radiology reports that have a follow-up recommendation using natural language processing (NLP) techniques and machine learning (ML) models.

Due to the nature of the radiology reports that serve as our input data, our approach makes extensive use of natural language processing (NLP). Although usage of structured reporting is certainly on the rise, the bulk of a radiology report's content may still be free text clinical narrative [5]. At our institution, the majority of a typical radiology report is still loosely structured text. Structured data must first be generated from the free text in order to perform analyses. Natural language processing allows for processing of large amounts of text that would not be possible using manual efforts.

NLP has many applications in radiology. Some applications that have been explored include diagnostic surveillance, cohort building for epidemiological studies, quality assessment, and clinical support services [6]. As radiology reports become more standardized, NLP techniques that use radiology reports as input will only become more reliable and consistent. Other studies have reported success in identifying recommendation statements [1, 2, 7, 8]. NLP techniques vary, but some fundamental procedures exist in most NLP pipelines. Generally, free text is first preprocessed, in an attempt to reduce the text to its fundamental semantic content. Then specific quantitative information, or features, are calculated from the processed text, which is used to train machine learning algorithms. This process is heavily dependent on the goals of the pipeline.

Methods

Dataset

The data set used in this study comprises reports for 6000 randomly sampled abdominal MRI, CT, and ultrasound examinations performed in 2016 and 2017 at one of the three hospitals in our urban health system. All these reports contained a numeric label assigned to the liver, pancreas, kidneys, and adrenal glands that reflects the degree of suspicion for malignancy in focal masses affecting these organs [9]. The labels categorize focal masses as benign, indeterminate, suspicious, or known malignancy. Patients with indeterminate lesions are expected to get follow-up imaging, while those with suspicious lesions are expected to get a biopsy or surgical resection. For our analysis, reports containing an indeterminate or suspicious label were considered as requiring follow-up, and all other studies were considered as not requiring follow-up. The template containing the numeric labels was stripped from all reports before our analysis. We used an 80:20 training set to test set ratio where 1200 radiology reports were randomly selected to be in the test set, and the remaining 4800 radiology reports were used for training (Table 1).

Classification Problem Formulation

To more rigorously formulate the problem, we create a binary classification function, which converts a free text report to a binary output. The classifier takes an input $X = [x_1, \ldots, x_k]$, a sequence of word tokens, and outputs a probability, from 0 to 1, that the input contains a follow-up recommendation. Depending on the preference for sensitivity or specificity, a

Table 1 Number of reports by modality and follow-up recommendation

US	2646
Follow-up indicated (+)	249
No follow-up indicated (−)	2397
Prevalence	9.4%
CT	2557
Follow-up indicated (+)	324
No follow-up indicated (−)	2233
Prevalence	12.7%
MR	797
Follow-up indicated (+)	162
No follow-up indicated (−)	635
Prevalence	20.3%
All modalities	6000
Follow-up indicated (+)	735
No follow-up indicated (−)	5265
Prevalence	12.3%

cutoff value is chosen such that any probability higher than the cutoff value is classified as containing a follow-up recommendation and any probability lower than the cutoff value is classified as not containing a follow-up recommendation. The pipeline uses the Python programming language (version 3.6.2) and the Natural Language Toolkit package (version 3.2).

Text Processing

The first portion of the pipeline involves using NLP to generate features for the machine learning algorithm. We first preprocess reports by converting all report text to lower case and removing all punctuation, symbols, and numbers. We then perform tokenization where each word is segmented, so that each report is represented by a list of words. Stop words, commonly used words that do not provide semantic value such as "the" and "a," were removed. Each remaining word was then reduced to its word stem using the Porter stemming algorithm [10], a commonly used stemmer. Some English words are ambiguous in their word stem, such as the word "axes" ambiguously being the plural of either "ax" or "axis," but these cases rarely arose in our data set.

As an example, the following report text:

> "Recommend contrast-enhanced MRI for further characterization.
> ATTENDING RADIOLOGIST AGREEMENT: I have personally reviewed the images and agree with the preliminary report without modification."

becomes the following after going through the preprocessing pipeline:

> "recommend contrast enhanc mri further character attend radiologist agreement have person review imag agre with preliminari report without modif"

Feature Engineering

We then generated feature vectors using the bag-of-words model. In this model, the term frequency of n-grams constitutes the features. We adjusted the number of unigrams, bigrams, and trigrams, and explored parameters for optimal performance. In general, the bag-of-words model creates a large number of features as the number of words in the vocabulary of all reports greatly exceeds the number of reports. If including n-grams with $n > 1$, then this potentially adds all combinations of n words. The number of features, n, greatly exceeds the number of input reports, N. Hence, features must be reduced in order to avoid overfitting of data, which is always a concern when $n > N$. Large feature spaces also increase computation time. In our approach, we first filter out all n-grams which occur in fewer than 10 reports to increase the generalizability of the model. During feature engineering, we found monograms, bigrams, and trigrams, to be most useful for model performance. Most n-grams with $n > 3$ occurred in fewer than 10 reports, and so would not be generalizable to new radiology reports.

We then further reduce the dimension of the feature space by preserving the most informative features determined by joint mutual information [11]. Bigrams and trigrams were often the most informative features and allowed for important context to be used by the machine learning algorithms. For example, the model learned that follow-up imaging recommendations often were at the conclusion of the report, followed by an attending radiologist agreement. The model associates the trigram "character attend radiologist" derived from "… recommended for further characterization. ATTENDING RADIOLOGIST AGREEMENT" with follow-up imaging. However, the bigram "further characterization" by itself was not predictive as many lesions were described as "not further characterized" and did not need follow-up. The position of the word "characterization" at the end of the report was important information conveyed by the trigram. The selection of the 1500 most informative features yielded the highest F-scores. Using more features increased runtime of training models without increasing F-scores.

Machine Learning Classification

Once features were generated, we compare three ML algorithms: naïve Bayes, decision tree, and maximum entropy (also referred to as the log-linear model). The NLTK implementation of each algorithm was used. The naïve Bayes classifier which independently assesses each feature does not have modifiable parameters nor randomized initial values, so the model generated for a given data set is identical each time. For the maximum entropy model, we employed the MEGA Model Optimization package [12] for parameter optimization and used 10 as the maximum number of iterations, as greater than 10 iterations showed marginal improvements. The decision tree uses the ID3 algorithm which optimizes for information gain. Parameters were set as 100 for maximum depth and 0.05 as the entropy cutoff. Increasing the depth parameter further did not improve performance. Finally, performance between the models was compared using accuracy and F-score of the algorithm on the testing data set.

Results

Of the 6000 reports used in the study, 735 reports contained follow-up recommendations, representing a 12.3% prevalence. The reports were randomized into training and testing sets in an 80:20 ratio. The training set contained 595 reports with a follow-up recommendation and 4205 without. The testing set contained 140 reports with a follow-up recommendation and 1060 without.

After varying the number of features and preprocessing steps, it was found that the best *F*-scores were achieved using tokenization and Porter stemming. In terms of feature selection, many trigrams were found to have high mutual information, and so we included unigrams, bigrams, and trigrams in the bag-of-words model. The top 10 most informative features are shown in Table 2.

With the aforementioned feature engineering techniques applied, the naive Bayes, decision tree, and maximum entropy models achieved accuracies of 74.5%, 86.1%, and 81.0% respectively. Table 3 displays the accuracy, F1 score, sensitivity, positive predictive value (PPV), specificity, and negative predictive value (NPV) of each ML model. The best overall performing model was found to be the decision tree model. Table 4 displays the full confusion matrix of the decision tree model.

Discussion

Our results show that in the follow-up recommendation detection task, performance of the classifier is dependent on NLP processing techniques, feature selection, and choice of ML algorithm. There exist increasingly advanced NLP processing techniques and ML models, such as global vectors for word representation [13], word2vec for creation of word embedding features, recurrent neural networks, and convolutional neural networks [14]. Nonetheless, this study focuses on traditional NLP techniques and adds to the current literature by assessing multiple traditional ML models simultaneously on a broad set of abdominal radiology reports. Moreover, there is a lack of literature comparing different ML algorithms in this specific classification task. Our findings agree with existing literature in NLP radiology applications that overall performance is dependent on combined optimization of NLP and ML algorithms [15–17]. In particular, selection of the most informative unigrams, bigrams, and trigrams with the decision tree model resulted in the highest performance in this task.

Table 2 Most informative features in training set

Feature	Follow-up indication ratio
renal neoplasm	51.9:1.0
evalu with enhanc	38.4:1.0
character attend radiologist	33.9:1.0
treatment zone measur	29.3:1.0
mri month	25.7:1.0
better evalu with	24.8:1.0
thi singl phase	24.8:1.0
measur most like	23.0:1.0
zone measur	23.0:1.0
malign not exclud	23.0:1.0

The decision tree model has a recall of 50% and a precision of 42%. The low precision compared to NPV is partially expected considering the low prevalence of follow-up indications in the data set. While perhaps not robust enough yet for clinical usage, this study demonstrates proof of concept and underlines the strength of the ML decision tree algorithm. Decision trees are well suited to tasks in which hierarchical categorical distinctions can be made [18]. We postulate that the decision tree model may have increased performance in the follow-up detection task over the other algorithms. Follow-up imaging is indicated in a number of scenarios and our data set contains a variety of imaging modalities and organs. This multitude of scenarios leading to a follow-up imaging recommendation may be better parsed by a decision tree classifier. In comparison, a naive Bayes classifier will consider each *n*-gram feature independently and thus will miss information related to the clinical scenario which might be represented as combinations of features. Maximum entropy models are theoretically appropriate for sparse feature sets, such as those generated by NLP techniques, wherein the number of features often exceeds the number of training inputs ($n > N$). We have, however, limited our feature set to 1500 features which offered improved performance in all models.

In our specific problem, follow-up imaging is indicated in ambiguous situations where malignancy can neither be ruled out nor confirmed. In contrast to typical diagnostic surveillance, where a model may be designed to detect a specific diagnosis, follow-up detection is inherently challenging as the indication for follow-up imaging may be clinically ambiguous. In essence, the algorithm must recognize when the likelihood of malignancy is neither too high to nor too low. The need to recognize two boundary conditions, as opposed to just one, like in many other diagnostic surveillance tasks, makes follow-up recommendation detection challenging. Other studies have successfully extracted recommendation sentences [1, 2, 7, 8], which may contain recommendations for follow-up, but this approach detects the radiologist's literal advice, such as "follow-up CT is recommended in three months." Thus, this approach is not a classification task with true-negative results and also does not address reports requiring follow-up without a specific follow-up recommendation sentence,

Table 3 Performance measure for each of the machine learning models after optimal preprocessing and feature engineering

	Accuracy (%)	F1 score (%)	Sensitivity (%)	PPV (%)	Specificity (%)	NPV (%)
Naïve Bayes	74.5	38.1	67.1	26.6	75.5	94.6
Decision tree	86.2	45.8	50.0	42.2	90.9	93.2
Maximum entropy	81.2	38.7	50.7	31.3	85.3	93.0

certainly a scenario that could lead to missed follow-up. NLP and machine learning algorithms have shown utility in various tasks, but there is a lack of literature comparing different ML algorithms in this specific task.

With regard to feature engineering, we have noted that the inclusion of informative unigrams, bigrams, and trigrams are beneficial to performance. Table 2 shows the most informative features as determined by joint mutual information. Unsurprising are *n*-grams such as "malign not exclud" and "better evalu with." Certain clinical features such as "renal neoplasm" are also highly correlated with an indication for follow-up imaging. This suggests that clinical semantic analysis using lexicons may be a useful generator of features. Semantic knowledge has been shown to be useful in other diagnostic surveillance tasks [19]. Though there is a lack of specific clinical indications for follow-up imaging, additional semantic processing could further separate institutional or individual radiologist's stylistic preferences from the underlying semantic meaning.

Further analysis of informative features revealed *n*-grams that indicate uncertainty, such as "lesion incomplet," "indetermin lesion," and "not exclud." More direct attempts at determining uncertainty may be beneficial in future pursuits. Intuitively, a radiology report that has high uncertainty will warrant follow-up imaging to gain further information on a lesion. Uncertainty, however, while highly prevalent in reports requiring follow-up, is also prevalent in radiology reports that do not require follow-up. Integrating uncertainty metrics as features may yield performance benefits and other recent studies have suggested methods for quantifying uncertainty [20].

To address limitations in our study, our data set, while sizable, was also broad. It contained 3 imaging modalities, 4 organ systems, multiple radiologists, and a low prevalence of follow-up imaging. This resulted in 735 total radiology reports with follow-up recommendations. With 20% of reports being set aside for testing, around 600 reports with follow-up indications were left over for the algorithm to learn from. Moreover, all reports came from our tertiary care center and therefore this study lacks inter-institutional validation. Inter-institutional reporting variation is a well-documented concern in applying NLP algorithms to radiology reports [6]. For example, our model has learned features that are not generalizable to all institutions. The *n*-grams representing attending attestations allow the model to infer which words are at the conclusion of a report. For institutions without attending attestations, our model will be unable to use those features.

With regard to report language style, our institution mandates a coding scheme to categorize lesions in abdominal imaging which may make our abdominal radiology reports non-generalizable. With knowledge that a code must be assigned, radiologists may be more mindful in characterizing lesions or radiologists may shorten the free text since there will be a numerical code with the report. Since the system was introduced in 2013, our radiologists are aware that this is an addition to the interpretation that they would otherwise issue. Our goal has never been to change the way radiologists write their reports, and we have repeatedly emphasized that with the users of the system. The code is used just for machine readability, and it is commonly understood by radiologists that the report should contain standard text conveying the recommendation in addition to the code. Part of the motivation behind the use of the templates is to enable automated identification and tracking of patients with indeterminate or suspicious findings.

Another limitation of this study is that ground truth is determined by the assigned numerical category, which is not infallible. A fair comparison of performance would involve review of free text reports without the associated imaging by radiologists, as the algorithm only analyzes the free text report. Nonetheless, this limitation highlights the potential benefit of using NLP algorithms as an adjunct to structured reports. Structured reporting allows for fast and simple access to follow-up recommendations but comes at the cost of

Table 4 Confusion matrix for the decision tree model predictions on the test set

		Predicted label	
		Follow-up indicated	No follow-up indicated
True label	Follow-up indicated	70	70
	No follow-up indicated	96	964

additional effort by radiologists, and may be coded incorrectly or not coded at all. Automated detection does not require additional effort by radiologists, but is less accurate. Institutions implementing structured reporting can use NLP algorithms to help generate structured fields or validate existing structured fields. If implemented in real time, such a system could create an alert that a report has a high probability of requiring follow-up imaging. This could be done while the radiologist is dictating the report and increase the compliance with follow-up recommendation protocols.

Conclusion

Our conclusions must be interpreted with the aforementioned limitations in mind. The decision tree model shows unexpected robustness compared to naive Bayes and maximum entropy models. Careful selection of features that have predictive power was also critical to performance. Follow-up recommendation detection is a challenging task that can certainly be addressed by explicit structured reporting of follow-up recommendations, but until a standardized system of doing so becomes prevalent in radiology, NLP and ML powered automated detection algorithms may assist in tracking the many patients who are at risk for adverse events due to delayed or missed follow-up imaging. While it is a complex task that is dependent on factors such as the institution, imaging modality, and individual radiologist's language, careful feature engineering and appropriate selection of machine learning algorithms are important strategies to improve the performance of follow-up detection classifiers.

References

1. Dutta S, Long WJ, Brown DF, Reisner AT: Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. Ann Emerg Med 62(2):162–169, 2013
2. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH: Automatic identification of critical follow-up recommendation sentences in radiology reports. AMIA Annu Symp Proc 2011:1593–1602, 2011
3. Cook TS, Lalevic D, Sloan C, Chadalavada SC, Langlotz CP, Schnall MD, Zafar HM: Implementation of an automated radiology recommendation-tracking engine for abdominal imaging findings of possible cancer. J Am Coll Radiol. 14(5):629–636, 2017
4. Langlotz CP: Structured radiology reporting: are we there yet? Radiology. 253(1):23–25, 2009
5. Bosmans JM, Peremans L, Menni M, De Schepper AM, Duyck PO, Parizel PM: Structured reporting: if, why, when, how-and at what expense? Results of a focus group meeting of radiology professionals from eight countries. Insights Imaging. 3(3):295–302, 2012
6. Pons E, Braun LM, Hunink MG, Kors JA: Natural language processing in radiology: a systematic review. Radiology. 279(2):329–343, 2016
7. Xu Y, Tsujii J, Chang EI: Named entity recognition of follow-up and time information in 20,000 radiology reports. J Am Med Inform Assoc. 19(5):792–799, 2012
8. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH: A text processing pipeline to extract recommendations from radiology reports. J Biomed Inform 46(2):354–362, 2013
9. Zafar HM, Chadalavada SC, Kahn CE, Cook TS, Sloan CE, Lalevic D et al.: Code abdomen: an assessment coding scheme for abdominal imaging findings possibly representing cancer. J Am Coll Radiol JACR. 12(9):947–950, 2015
10. The Porter Stemming Algorithm. https://tartarus.org/martin/PorterStemmer/. Accessed June 1, 2017.
11. Peng HC, Long F, Ding C: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27(8):1226–1238, 2005
12. MEGA Model Optimization Package. http://legacydirs.umiacs.umd.edu/~hal/megam/version0_3/. Accessed June 1, 2017.
13. Pennington J, Socher R, Manning C: Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543
14. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, Langlotz CP, Amrhein TJ, Lungren MP: Deep learning to classify radiology free-text reports. Radiology. 286(3):845–852, 2018
15. Chen PH, Zafar H, Galperin-Aizenberg M, Cook T: Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. J Digit Imaging. 31(2):178–184, 2018
16. Hassanpour S, Langlotz CP, Amrhein TJ, Befera NT, Lungren MP: Performance of a machine learning classifier of knee MRI reports in two large academic radiology practices: a tool to estimate diagnostic yield. AJR Am J Roentgenol. 208(4):750–753, 2017
17. Hassanpour S, Bay G, Langlotz CP: Characterization of change and significance for clinical findings in radiology reports through natural language processing. J Digit Imaging. 30(3):314–322, 2017
18. Bird, Steven, Ewan, Klein, and Loper, Edward (2009), Natural language processing with Python, O'Reilly Media..
19. Garla V, Taylor C, Brandt C: Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. J Biomed Inform 46(5):869–875, 2013
20. Reiner BI: Quantitative analysis of uncertainty in medical reporting: creating a standardized and objective methodology. J Digit Imaging. 31(2):145–149, 2018