SIIM 2017 Scientific Session
Analytics & Deep Learning Part 1
Thursday, June 1 | 1:15 pm – 2:45 pm

Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports

Po-Hao Chen, MD, MBA, Hospital of the University of Pennsylvania; Hanna Zafar; Tessa S. Cook

Hypothesis

Artificial intelligence software's ability to predict radiologist intent in an oncologic diagnostic report relies on the co-dependent, combinatorial optimization of both the natural language processing and machine learning algorithms.

Background

The advent of structured reporting may improve the availability of standardized data elements in a radiology report for text mining. However, most radiology reports remain unstructured. For named-entity recognition, regular-expression and search-based report analytics have been shown to extract specific critical diagnoses successfully (1,2). Natural language processing (NLP) js increasingly being used to analyze radiology reports for oncologic imaging (3,4). For instance, the presence of specific malignant diagnoses such as lung cancer and colon cancer have been previously examined (5).

While some emergent diagnoses can be reported with certainty, in many cases such as oncologic follow-up, the reporting language may be less clear, reflecting the inherent uncertainty in such radiologic evaluations. In this study, we assess the effect of multiple NLP techniques and machine learning algorithms on the automatic detection of the radiologist's intent in oncologic evaluations.

Methods

At the authors' institution, all cancer follow-up CT and MRI examinations received an assigned score as part of the formal diagnostic report, termed "Code Oncology", adapted from a previously published initiative for reporting focal abdominal lesions (6). The interpreting radiologist was required to assign a value to each of the two specified categories: (a) interval evolution of existing lesions and (b) interval appearance of new lesions. The "existing lesion" category has eight possible values: No previously documented cancer, complete response, significant improvement, mild improvement, stable, mild progression, significant progression, mixed response, and indeterminate. The "new lesion" category contains three possible values: no new lesion, possible new lesion, and definite new lesion. The codes were designed for clinical use and assigned at the discretion of the interpreting board-certified radiologist.

Between 4/1/2015 and 11/1/2016, a total of 9,418 cross-sectional abdomen/pelvis CT and MRI were performed which contain the manually created Code Oncology performed cancer follow-up were included in the initial analysis. Reports including the "mixed response," "indeterminant," or "possible new lesion" categories were excluded from the study due to the wide variation in practice for these assignments.

We created four labels for overall assessment. "Progression" was defined as either interval development of new lesion(s) or either mild or significant progression of existing lesions. "Improvement" was defined as no interval development of a new lesion and either mild or significant improvement of existing lesions. "Stable disease" was defined as no interval development of a new lesion and stable appearance of existing lesions. "Resolution/no cancer" was defined as the absence of any new lesion and either "no previously documented cancer" or "complete response."

The structured report text was parsed and then removed from the report text prior to report pre-processing. Pre-processing was performed using the Azure Machine Learning Studio, using a combination of the Python programming language, the Natural Language Toolkit Python package, and native preprocessing modules (7). Text header detection was performed using regular expressions to segment the radiology report by section. Only the impression was utilized, as the use of impression yielded more accurate performance compared to using both findings and impression based on authors' prior experience. If more than one impression bullet point existed, then the impression was included both in total as well as separated by each bullet point. All report text was then converted to lower case and all punctuations removed. For each section, evaluation was performed after applying an English word tokenizer both with and without stop word removal and both with and without applying a Porter stemmer.

Three forms of text feature vectorization using the bag-of-words model were compared: term frequency-inverse document frequency weighting (TF-IDF), term frequency weighting (TF), and 16-bit feature hashing. Vectorization parameters were adjusted for the overall best predictive performance defined by the ML model's micro-average F-measure (8). Parameters adjusted include N-gram (up to five-gram). For TF and TF-IDF, K-skip size, minimum N-gram document absolute frequency, and maximum N-gram document ratio were also explored for optimal performance. Filter-based feature selection was performed to select the most relevant features using mutual Information (9).

Five machine learning algorithms were compared in the present study, including logistic regression (LR), random decision forest (RDF), one-vs-all support vector machine (SVM), one-vs-all Bayes point machine (BPM), and fully-connected neural network (NN). Input data was stratified by classification label and randomly assigned into training (70%) or testing (30%) datasets. The Bayes point machine was implemented to train for 60 iterations with bias. The training data was divided into 5 folds to perform an 8-run random sweep with cross-validated hyperparameter model tuning to identify the best parameter set for each of the remaining four machine learning algorithms. The performance was measured using a micro-average F-measure and average classification accuracy using the testing dataset (8,10).

Results

Of the 9418 examinations performed within the study timeframe, 8614 examinations met the inclusion criteria. Of these, 2800 were manually categorized as "resolution/no cancer", 2498 categorized as "progression," 2132 categorized as "stable disease" and 1184 categorized as "improvement."

The set of NLP techniques which yielded the best predictive accuracy and F-measure is referred hereafter as "reference NLP techniques" consisting of tokenized unigrams and bigrams with term frequency – inverse document frequency (TF-IDF), stop word removal, Porter stemming, and filter-based feature selection limited to the top 1000 features. Using the reference NLP techniques on the testing dataset, the Bayes point machine algorithm achieved an 89.5% average classification accuracy. After hyperparameter model tuning, the best performing multi-class logistic regression algorithm, random decision forest algorithm, fully-connected neural network, and support vector machine achieved an average predictive accuracy of 90.2%, 90.0%, 88.3%, and 90.6%, respectively. Table 1 displays the results from training and testing accuracy as well as F-measures.

Table 1

	Training		Testing	
	Accuracy	**F-Measure**	**Accuracy**	**F-Measure**
Bayes Point Machine	91.5%	0.830	89.5%	0.791
Logistic Regression	91.5%	0.829	90.2%	0.803
Random Decision Forest	98.1%	0.962	90.0%	0.800
Neural Network	91.4%	0.829	88.3%	0.765
Support Vector Machine	91.4%	0.828	90.6%	0.813

Table 1 – Average multi-class classification accuracy and F-measure for each of the 5 trained machine learning models utilizing the optimal parameters after hyperparameter tuning.

With other elements of the reference NLP techniques held constant, stop word removal (SWR) slightly improved the micro-average F-measure for all ML algorithms relative to no SWR. Word stemming slightly improved the performance of BPM, NN, and SVM, but did not impact or minimally degraded the F-measure of LR and RDF. TF-IDF was superior to TF alone for BPM, NN, and SVM but slightly decreased accuracy in RDF and had no effect in LR. Using feature hashing rather than TF-IDF improved the runtime of model training but decreased micro-average F-measure for BPM, LR, and SVM, with minimal performance effect on RDF and NN. Table 2 demonstrates the relative contribution of each of the NLP parameters.

Table 2

	Bayes Point Machine	**Logistic Regression**	**Random Decision Forest**	**Neural Network**	**Support Vector Machine**
Reference	0.791	0.803	0.800	0.765	0.813
No SWR	+0.003	-0.009	-0.002	-0.005	-0.013
No word stemming	-0.002	+0.004	+0.0004	-0.005	-0.003
TF	-0.001	0.000	+0.004	-0.005	-0.003
Feature Hash	-0.007	-0.016	0.000	+0.001	-0.019

Table 2 – Effect of NLP parameters on micro-averaged F-measure score. Reference - stop-word removal, application of Porter word stemmer, with feature extraction using unigram and bigrams, term frequency - inverse document frequency (TF-IDF) weighting, top 1000 features by mutual information (MI) filter selection. SWR – Stop Word Removal. TF – term frequency.

A combination of unigrams and bigrams outperform other lengths of contiguous word series for all ML algorithms except for RDF, which performed best with a combination of unigrams, bigrams, as well as trigrams (Figure 1).

Figure 1

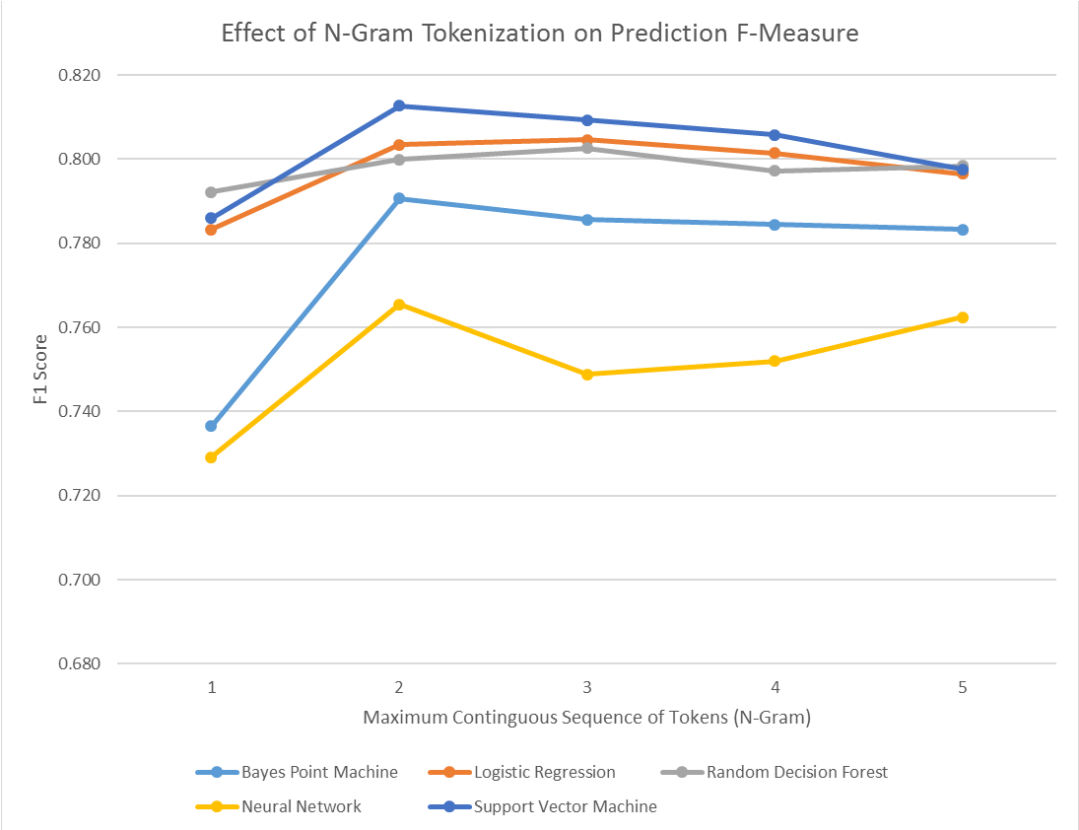

Table 3 lists the top 15 most discriminating word features ranked by mutual information. While LR and NN performed best with all the N-gram features, the other ML algorithms performed best when only the top 1000 features are used based on the filter-based selection. The effect of filter-based feature selection on F-measure of all 5 ML algorithms is shown in Figure 2.

Figure 2

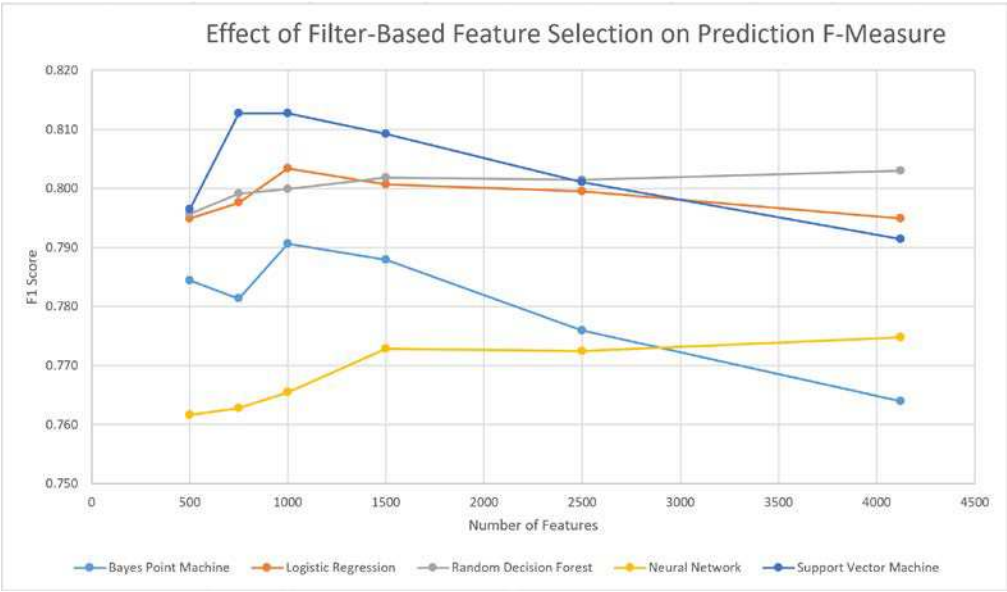

Table 3

Feature	Mutual Information
decreas	0.143
increas	0.140
abdomen	0.126
progress	0.117
decreas size	0.116
abdomen pelvi	0.114
size	0.110
pelvi	0.109
new	0.105
increas size	0.100
recurr metastat	0.095
interv	0.095
metastasi	0.092
stabl	0.086
recurr	0.085

Table 3 – Top 15 most differentiating features after applying unigram and bigram tokenization, term frequency-inverse document frequency, and Porter stemmer.

Discussion

The present study uses standardized reporting structures embedded within formal diagnostic reports as the ground truth for machine learning. Our results show that the performance of radiology report classification is likely dependent on both the machine learning algorithm and on the natural language processing parameters. The data adds to current literature by assessing multiple machine learning algorithms simultaneously. Our findings agree with existing literature in electronic report text mining that support vector machine (SVM) performs well in classification tasks (5).

The best predictive performance was achieved using SVM with reference NLP parameters. The present study further assessed the effect of optimizing NLP parameters by assessing the impact of each modification on five different ML algorithms. Stop word removal generally improves the F-measures of all ML algorithms except for Bayes point machine (BPM). The use of TF-IDF rather than TF alone had a modest to equivocal effect on F-measures. The use of 16-bit feature hashing significantly improved the runtime of all five algorithms but decreased the F-measure of BPM, LR, and SVM while having minimal or no impact on random decision forest (RDF) and fully-connected neural network (NN) algorithms.

The relative performance of SVM decreases with more features included. Specifically, with greater than 2500 text features, RDF outperforms SVM when other parameters are held constant. When the full set of 4122 text features are used, both RDF and logistic regression (LR) performed better than SVM. The interval decrease in performance in SVM, LR, and BPM – but not RDF or NN – as the size of the feature set increases is likely related to overfitting. Specifically, our findings agree with existing literature that RDS can be relatively resistant to performance penalties from overfitting (11,12).

Future direction of this project includes the application of additional natural language processing algorithms. For instance, convoluted neural networks (CNN) have shown remarkable success in image recognition and classification, and have been applied to NLP feature extraction in medical literature such as semantic models (13,14). Additionally, the use of skip-gram models in the future may yield improved performance over TF, TF-IDF, and hashing mechanics. The present study is limited by the size of its annotated dataset, as models like word2vec rely on significantly larger training sets for accurate representation.

Conclusion

Although natural language processing and machine learning algorithms have the potential to accurately classify the radiologist's diagnostic intent in the oncologic interpretation, the overall performance depends on the combinatorial optimization of both the NLP and ML algorithms.

References

1. Lakhani P, Kim W, Langlotz CP. Automated extraction of critical test values and communications from unstructured radiology reports: an analysis of 9.3 million reports from 1990 to 2011. Radiology. 2012 Dec;265(3):809–18.
2. Lakhani P, Kim W, Langlotz CP. Automated Detection of Critical Results in Radiology Reports. J Digit Imaging. 2012 Feb;25(1):30–6.
3. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. Radiogr Rev Publ Radiol Soc N Am Inc. 2016 Feb;36(1):176–91.
4. Yim W-W, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. JAMA Oncol. 2016 Jun 1;2(6):797–804.
5. Kocbek S, Cavedon L, Martinez D, Bain C, Mac Manus C, Haffari G, et al. Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources. J Biomed Inform. 2016 Oct 11;
6. 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. 2015 Sep;12(9):947–50.
7. Bird S, Klein E, Loper E. Natural language processing with Python. 1st ed. Beijing ; Cambridge [Mass.]: O'Reilly; 2009. 479 p.
8. Lipton ZC, Elkan C, Naryanaswamy B. Optimal Thresholding of Classifiers to Maximize F1 Measure. Mach Learn Knowl Discov Databases Eur Conf ECML PKDD Proc ECML PKDD Conf. 2014;8725:225–39.
9. Bennasar M, Hicks Y, Setchi R. Feature selection using Joint Mutual Information Maximisation. Expert Syst Appl. 2015 Dec;42(22):8520–32.
10. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc JAMIA. 2005 Jun;12(3):296–8.
11. Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R, et al. Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: assessing chance correlation and prediction confidence. Environ Health Perspect. 2004 Nov;112(16):1622–7.
12. Liu X, Song M, Tao D, Liu Z, Zhang L, Chen C, et al. Random forest construction with robust semisupervised node splitting. IEEE Trans Image Process Publ IEEE Signal Process Soc. 2015 Jan;24(1):471–83.
13. Wang J, Zhang J, An Y, Lin H, Yang Z, Zhang Y, et al. Biomedical event trigger detection by dependency-based word embedding. BMC Med Genomics. 2016 Aug 10;9 Suppl 2:45.

14. Wei W, Marmor R, Singh S, Wang S, Demner-Fushman D, Kuo T-T, et al. Finding Related Publications: Extending the Set of Terms Used to Assess Article Similarity. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2016;2016:225–34.

Keywords

machine learning, natural language processing, support vector machine, bayes point machine, neural network, logistic regression, random decision forest, TF-IDF