Radiology

Natural Language–based Machine Learning Models for the Annotation of Clinical Radiology Reports[1]

John Zech, MA
Margaret Pain, MD
Joseph Titano, MD
Marcus Badgeley, MEng
Javin Schefflein, MD
Andres Su, MD
Anthony Costa, PhD
Joshua Bederson, MD
Joseph Lehar, PhD
Eric Karl Oermann, MD

Purpose: To compare different methods for generating features from radiology reports and to develop a method to automatically identify findings in these reports.

Materials and Methods: In this study, 96303 head computed tomography (CT) reports were obtained. The linguistic complexity of these reports was compared with that of alternative corpora. Head CT reports were preprocessed, and machine-analyzable features were constructed by using bag-of-words (BOW), word embedding, and Latent Dirichlet allocation–based approaches. Ultimately, 1004 head CT reports were manually labeled for findings of interest by physicians, and a subset of these were deemed critical findings. Lasso logistic regression was used to train models for physician-assigned labels on 602 of 1004 head CT reports (60%) using the constructed features, and the performance of these models was validated on a held-out 402 of 1004 reports (40%). Models were scored by area under the receiver operating characteristic curve (AUC), and aggregate AUC statistics were reported for *(a)* all labels, *(b)* critical labels, and *(c)* the presence of any critical finding in a report. Sensitivity, specificity, accuracy, and F1 score were reported for the best performing model's *(a)* predictions of all labels and *(b)* identification of reports containing critical findings.

Results: The best-performing model (BOW with unigrams, bigrams, and trigrams plus average word embeddings vector) had a held-out AUC of 0.966 for identifying the presence of any critical head CT finding and an average 0.957 AUC across all head CT findings. Sensitivity and specificity for identifying the presence of any critical finding were 92.59% (175 of 189) and 89.67% (191 of 213), respectively. Average sensitivity and specificity across all findings were 90.25% (1898 of 2103) and 91.72% (18351 of 20007), respectively. Simpler BOW methods achieved results competitive with those of more sophisticated approaches, with an average AUC for presence of any critical finding of 0.951 for unigram BOW versus 0.966 for the best-performing model. The Yule I of the head CT corpus was 34, markedly lower than that of the Reuters corpus (at 103) or I2B2 discharge summaries (at 271), indicating lower linguistic complexity.

Conclusion: Automated methods can be used to identify findings in radiology reports. The success of this approach benefits from the standardized language of these reports. With this method, a large labeled corpus can be generated for applications such as deep learning.

Online supplemental material is available for this article.

[1] From the Departments of Radiology (J.Z., J.T., J.S., A.S.) and Neurosurgery (M.P., M.B., A.C., J.B., E.K.O.), Icahn School of Medicine, 1 Gustave Levy Pl, New York, NY 10029; Verily Life Sciences, South San Francisco, Calif (M.B.); and Department of Bioengineering and Bioinformatics, Boston University, Boston, Mass (J.L.). Received May 16, 2017; revision requested June 30; revision received August 1; accepted September 16; final version accepted November 18. **Address correspondence to** E.K.O. (e-mail: *eric.oermann@mountsinai.org*).

Unstructured text notes contained within the electronic medical record (EMR) are recognized as a rich but difficult-to-access source of medical information. Addressing this has been a goal of clinical informatics for decades, and the development of natural language processing (NLP) algorithms to automatically extract structured information from the EMR is well described throughout the medical literature (1–7). Within the field of radiology, there has also been wide interest in developing NLP tools for epidemiologic cohort construction, quality assurance, clinical decision support, and other applications (8–12).

Machine learning describes a broad collection of techniques developed by computer scientists and statisticians that "focuses on the question of how to get computers to program themselves" (13). Using these techniques, powerful algorithms to make inferences from data can be learned automatically with minimal human input. *Deep learning* describes a particular subcategory of machine learning techniques that use multiple layers of neural networks to perform inference (14). Deep learning models with varied architectures have been applied successfully in many different domains, including image recognition (convolutional neural networks) and natural language processing (long short-term memory networks) (14). Deep learning–based image recognition techniques often require large amounts of training data to achieve high accuracy; for example, the standard ILSVRC competition data set, on which such algorithms are frequently trained and evaluated, includes 1.2 million training images (15). Deep learning–based approaches have increasingly been applied to medical image analysis, requiring large numbers of labeled medical images to facilitate these techniques (16). Many past efforts to automatically extract labels from radiology reports have utilized rule-based systems that were handcrafted to a specific corpus of reports (17). However, machine learning–based label extraction systems have become increasingly popular because of their scalability, ease of use, and rapidly improving accuracy (8,12).

In this study, we approach the problem of generating clinical labels for a large repository of radiology reports as a machine learning problem with an emphasis on scalability and generalizability. Over the past few years, there has been substantial progress within the machine learning community on performing NLP because of the availability of large data sets and high-performance computing. Despite progress in the general case, these newer, more advanced machine learning–based approaches have been only sparsely described within the medical literature (16,18–20). Our purpose was to compare different methods for generating features from radiology reports and to develop a method to automatically identify findings in these reports.

Implications for Patient Care

- Natural language processing approach allows the annotation of a large corpus of radiologic reports by using only a small labeled subset.
- This large labeled corpus facilitates the training of deep learning–based models to identify findings in imaging data.

Materials and Methods

Data Sets

Radiology reports.—Our study was approved by the Mount Sinai Institutional Review Board, and all data were stored locally on the hospital premises on a dedicated computing resource. Three authors (M.B., E.K.O., and J.L.) were employed by Verily Life Sciences (South San Francisco, California) at the time of this work, in addition to their academic affiliations. Verily Life Sciences did not provide financial support for this study and has no financial interest in it. The first author (J.Z.) had control of the data and material submitted for publication. Three corpora of radiology reports, consisting of all head computed tomographic (CT) scans, all hip radiographs, and all chest radiographs obtained at Mount Sinai Hospital and Mount Sinai Queens between 2010 and 2016, were collected for use in this study (Table 1). These three types of reports were selected as they were expected to contain markedly different language, and a large sample of each report was available for analysis. Corpora were assembled from cases stored within the hospital picture archiving and communication system. In total, 96303 head CT reports were available (Fig 1).

Consensus-based clinical entity label generation.—A subset of the head CT reports was selected to be labeled manually. Reports were randomly sampled from each year to generate a 1004-report corpus for annotation with reference-standard labels. Clinical entities were generated by utilizing the United Medical Language System Concept Unique Identifier and ordered into a taxonomy. Three physicians (two postgraduate year 4 radiology residents and a postgraduate year 4 neurosurgery

https://doi.org/10.1148/radiol.2018171093

Content code: IN

Radiology 2018; 287:570–580

Abbreviations:

AUC = area under the receiver operating characteristic curve
BOW = bag of words
DM = distributed memory
DV = document vector
EMR = electronic medical record
I2B2 UTHEALTH = Informatics for Integrating Biology and the Bedside 2014 De-identification and Heart Disease Risk Factors Challenge
LDA = Latent Dirichlet allocation
NLP = natural language processing
t-SNE = t-distributed stochastic neighbor embedding
TTR = type-to-token ratio

Author contributions:

Guarantors of integrity of entire study, J.Z., M.P., J.T., A.S., E.K.O.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, J.Z., M.P., J.T., M.B., J.S., A.S., J.B., J.L., E.K.O.; clinical studies, J.T., J.S., A.C.; experimental studies, J.Z., M.P., J.T., A.S., A.C., E.K.O.; statistical analysis, J.Z., J.T., J.L., E.K.O.; and manuscript editing, all authors

Conflicts of interest are listed at the end of this article.

Table 1

Analysis of Lexical Complexity of Two Common English Language Corpora (Reuters News and Gutenberg), Amazon Product Reviews, Hospital Discharge Summaries (I2B2), and Three Large Radiology Corpora (Head CT Scans, Hip Radiographs, Chest Radiographs)

Corpus	Average Normalized Relative Complexity
Reuters	1.00
Gutenberg	0.94
Hip radiographs	0.03
Head CT scans	0.08
Chest radiographs	0.04
I2B2 UTHEALTH	1.36
Amazon pet product reviews	1.01
Amazon cellular accessories reviews	0.91

resident) independently generated binary (true or false) labels for these reports that indicated the presence or absence of the specified clinical entities, and a consensus set of reference-standard labels was generated by majority vote (21). The Fleiss κ, a measure of interrater agreement, was calculated for each individual label, and only labels with κ of 0.60 or greater were considered for analysis. Whenever we discuss "labels" in this article, we are referring to these human-generated report annotations. In total, there were 55 labels included in the analysis, 20 of which were classified as critical labels on the basis of the criteria used at our hospital, which are derived from a previously published report (36). At our institution, reporting of acute ischemic stroke is required within 15 minutes of image acquisition, and reporting of intracranial hemorrhage, airway compromise, acute spinal cord compression, ruptured aneurysm, and markedly misplaced lines and tubes is mandated within 1 hour of image acquisition. Unexpected masses are reported within 8 hours of image acquisition. In addition, findings not contained within this specific list that may alter patient treatment acutely are reported urgently. An aggregate measure indicating whether a report contained any critical finding was determined to be true if any of the 20 critical findings were present, and false otherwise.

Comparative nonradiology text corpora.—Several alternative text corpora were gathered for the purposes of characterizing the lexical and semantic structure of radiology reports prior to analysis. These included subsets of the Reuters corpus of more than 10 000 news stories, the Gutenberg corpus of more than 3000 English language books, and the Informatics for Integrating Biology and the Bedside 2014 De-identification and Heart Disease Risk Factors Challenge (I2B2 UTHEALTH) corpus of discharge summaries (22–25). To compare with documents that were similarly constrained in their topic matter, we also analyzed Amazon product reviews from the "Cell Phones and Accessories" and "Pet Supplies" departments (26).

Complexity Analysis

We performed an initial lexical complexity analysis separately for head CT scans, chest radiographs, and hip radiographs, as well as for each of the comparative nonradiology text corpora. Because of differences in corpora size, we calculated complexity measures across entire corpora, as well as averaged measures across randomly sampled partitions of 500 000 tokens per partition. A token is the fundamental unit into which documents are subdivided in an NLP analysis: In a unigram analysis, all single words or other strings of characters separated by spaces or punctuation in reports constitute the set of tokens. To capture lexical complexity, we calculated simple metrics such as the number of unique words, the number of unique bigrams (two words joined and treated as a single token, eg, "acute_hemorrhage"), the type-token ratio (TTR), which divides unique word count by total word count, and the Yule I (which, similarly to TTR, increases when rare words occur more frequently in a document, but is calculated based on the distribution of the number of words that occur a given number of times) (27). A full description of these measures is given in Appendix E1 (online). We also trained simple first-order Markov models (which assume that the probability of a given word appearing in a sentence depends only on the prior word) on each corpus and calculated the average entropy rate for each model (roughly equivalent to the average per-word entropy, a measure that reflects the variety of language present, specifically defined as the average of the negative logarithm of the probability assigned by the model to each word appearing in a corpus) (28). Utilizing these features as measures of lexical complexity and utilizing the Reuters corpus as a baseline, we scaled each feature such that the score of the Reuters corpus was 1 and then took a simple, normalized average to create an aggregate measure of complexity relative to the Reuters corpus. This "average normalized relative complexity" measure equals 1 when the corpus under consideration is equally complex as the Reuters corpus, is between 0 and 1 for simpler corpora, and is greater than 1 for more complex corpora.

Preprocessing

All text was preprocessed in a standardized manner to facilitate machine learning techniques. All sections of each report, including any addenda made, were used in feature generation. Reports were segmented by using standard approaches based on white spacing and punctuation.

A generic set of stop words, several radiology-specific phrases (eg, "CT"), and other noninformative characters (eg, numbers, "_," "-," ",") were removed. Text was converted to lowercase, and words were stemmed by using the Porter stemming algorithm (29). When specified, *n*-grams were constructed based on the words in the documents (1-, 2-, or 3-grams). *N*-grams that occurred fewer than 20 times in training data were excluded from the analyses to ensure that the constructed vocabulary was meaningful to the corpus at large.

Featurization

We note the distinction between machine learning approaches that are "supervised" and those that are "unsupervised." Supervised learning is a subfield of machine learning concerned with predicting outcomes given sets of features (eg, regression, classification). In contrast, unsupervised learning is performed when there are no outcomes and our goal is instead to learn associations and patterns between various sets of features (eg, clustering, dimensionality reduction) (30). Three general unsupervised approaches to generating predictive features were compared. All approaches were trained exclusively on the 95 299 unlabeled reports and were then subsequently incorporated into supervised models. A standard grid search was performed for hyperparameter selection for all algorithms used during both featurization and subsequent modeling. Hyperparameters refer to algorithm parameters that cannot be directly learned during training and must be specified beforehand. As a benchmark for the more sophisticated methods, a simple bag of words (BOW) model was used. BOW discards grammar and context and exclusively utilizes document-level word occurrences as its features. For each document, a BOW vector with a length equal to the size of the vocabulary of the corpus was generated in which each entry corresponded to the presence of a specific word in the document (1 if the word was present in document, 0 otherwise). For example, in the unigram vocabulary of 3291 words, each document had a corresponding BOW unigram vector of length 3291 in which each entry in the vector corresponded to a specific word. A document that contained only the phrase "Impression: acute hemorrhage" would have a 1 at each of the three entries in the vector corresponding to "impression," "acute," and "hemorrhage" and a 0 at every other entry in the vector. This vector ("BOW-Unigram") was then used as input to predict each human-assigned label. This approach was extended to phrases of two words in "BOW-Bigram" (eg, "impression_acute," "acute_hemorrhage") and three words in "BOW-Trigram" (eg, "impression_acute_hemorrhage") to determine if capturing longer phrases could improve predictive accuracy.

Latent Dirichlet allocation (LDA) topic models were built on the unlabeled reports (31,32). This model was chosen because it represents the most popular unsupervised method of topic discovery based on probabilistic generative models, which specify a large

Figure 1

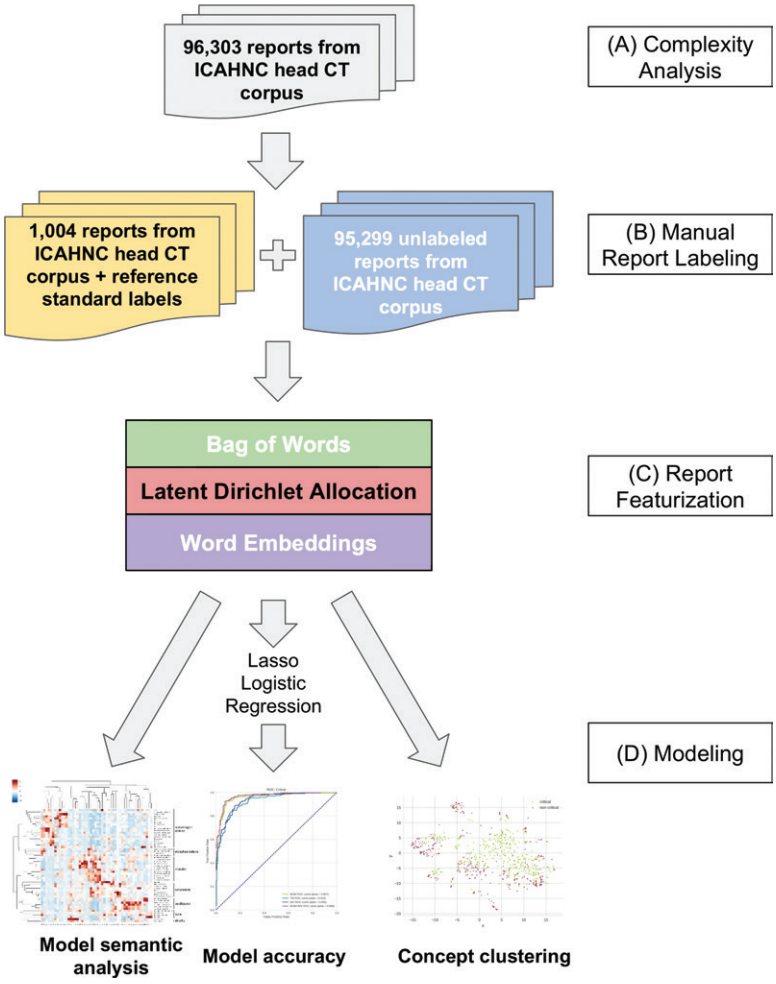


Figure 1: The overall workflow of the natural language processing analysis beginning with *(A)* calculating the complexity benchmarks, *(B)* obtaining reference-standard labels manually, *(C)* feature engineering with several different methods, and *(D)* evaluation of the final predictive models. ICAHNC = Mount Sinai Hospital and Mount Sinai Queens.

Table 2

Area under the Curve for the Best-performing Models within Each Machine Learning Feature Engineering and Model Category

	Overall (n = 55 findings)				Critical (n = 20 findings)				Critical Finding Present
Held-out AUC according to Model Type	AUC	Minimum	Maximum	SD	AUC	Minimum	Maximum	SD	AUC
BOW									
Uni-, bi-, and trigrams	0.954	0.816	1.000	0.041	0.953	0.857	0.997	0.040	0.957
Unigrams	0.950	0.814	1.000	0.046	0.951	0.857	0.998	0.040	0.951
Bigrams	0.894	0.515	1.000	0.099	0.912	0.729	0.978	0.061	0.937
Trigrams	0.790	0.500	0.992	0.133	0.824	0.533	0.961	0.116	0.853
DM-DV									
Document-embedding vector	0.761	0.500	0.932	0.096	0.831	0.695	0.911	0.066	0.876
Average word-embedding vector	0.917	0.746	0.992	0.056	0.935	0.746	0.991	0.055	0.935
Document-embedding vector + average word-embedding vector	0.917	0.746	0.993	0.056	0.934	0.746	0.991	0.055	0.935
LDA									
50 Topics, uni-, bi-, trigrams	0.820	0.500	0.995	0.109	0.887	0.730	0.987	0.074	0.852
50 Topics, unigrams	0.849	0.566	0.994	0.098	0.905	0.832	0.979	0.049	0.936
50 Topics, bigrams	0.763	0.454	0.987	0.154	0.877	0.714	0.983	0.089	0.846
50 Topics, trigrams	0.690	0.399	0.992	0.169	0.812	0.500	0.978	0.149	0.776
100 Topics, unigrams	0.860	0.500	0.994	0.118	0.922	0.747	0.989	0.060	0.929
200 Topics, unigrams	0.872	0.550	0.994	0.091	0.914	0.754	0.981	0.060	0.923
BOW (uni-, bi-, trigrams) + DM-DV average word-embedding vector	0.957	0.827	1.000	0.039	0.959	0.857	0.997	0.037	0.966

Note.—AUC = area under the receiver operating characteristic curve, BOW = bag of words, DM-DV = distributed memory document vector, LDA = Latent Dirichlet allocation, SD = standard deviation.

joint distribution over many related variables (33). Although the details of LDA are beyond the scope of this article, these models can be conceptualized by assuming that a document consists of words and topics. While we know the words in the document, we cannot directly observe the topics, and must infer them. LDA assumes that certain topics are more likely to be associated with certain words, and we can therefore use the words in the document to infer the topics. Topics can often be directly interpreted by humans as semantically meaningful (34). The learned topic assignments in each document can be understood as characterizing the document's semantics. We utilized these topic assignments as our LDA features. For an LDA model trained on head CT data, we reported the average document topic distribution by clinical label and the top five words associated with each topic (35). The number of topics in an LDA model is a hyperparameter that must be prespecified before model training, and we report results for LDA models with 50, 100, and 200 topics.

For our embedding-based approaches, we relied on a distributed memory (DM)-document vector (DV) model (19,32). We included this model because it is a state-of-the-art discriminative method that offers a fundamentally different alternative to LDA's probabilistic generative approach. DM-DV is an extension of a "continuous bag of words" model, an artificial neural network-based approach that utilizes a predictive task, in this case predicting a word given the words that precede and follow it, to learn an n-dimensional vector of numbers, called an embedding vector, that captures some of the semantic content of the underlying words. These learned embedding vectors can subsequently be utilized as features in predictive models or as a means of characterizing the underlying semantic space. DM-DV extends the skip-gram model by learning both word embedding vectors as well as an additional document or paragraph embedding vector (document in our case) that captures some of the semantic content of a given document or paragraph. While an active area of research, a straightforward approach for feature generation is to sum the word embedding vectors for each word in a document and then divide by the total number of words in that document to get an "average word embedding vector" for each document. We use these average word embedding vectors as features, as well as the individual document embedding vectors learned by the DM-DV model with an embedding dimensionality of 400 and a window of three. Therefore, for each document, these two vectors were calculated and were then used separately and in combination as input to predict each human-assigned label. We report examples of learned word embeddings, average word embedding vectors for critical versus noncritical findings, and

Figure 2

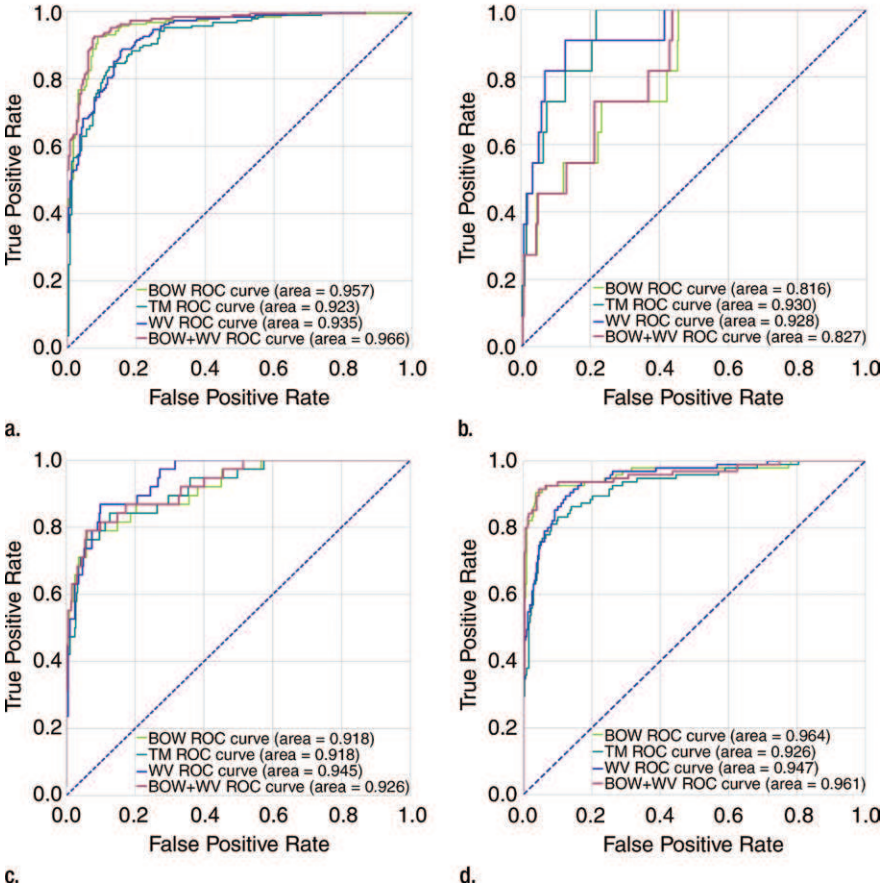


Figure 2: Graphs show ROC curves for the different models for several of the higher order labels for **(a)** presence of a critical finding, **(b)** fracture, **(c)** hemorrhage, and **(d)** stroke, infarction, or ischemia. BOW = bag of words, ROC =receiver operating characteristic, TM = Latent Dirichlet allocation topic model, WV = average word embedding vector.

Table 3

Performance Metrics for the Best Overall Model (BOW+DM-DV) on a Single Binary Prediction Task (Predicting Whether a Report Contains a Critical Result), as Well as Its Overall Performance Predicting All Labels

BOW+DM-DV	All Labels (n = 22 110)	Critical Finding Present (n = 402)
Sensitivity (%)	90.25 (1898/2103)	92.59 (175/189)
Specificity (%)	91.72 (18 351/20 007)	89.67 (191/213)
Accuracy (%)	91.58 (20 249/22 110)	91.04 (366/402)
F1	0.671 (3796/5657)	0.907 (350/386)

Note.—BOW+DM-DV = model using both bag of words and average word embedding vector features, F1 = (two times precision times recall)/(precision plus recall).

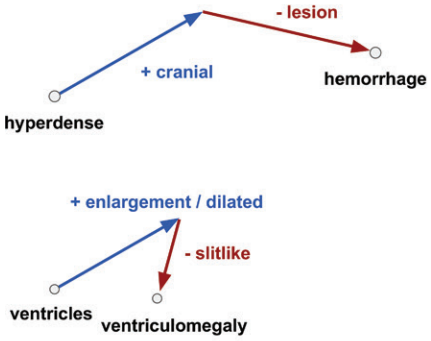
clusters among learned word embeddings in the head CT corpus (36). We used t-distributed stochastic neighbor embedding (t-SNE) to generate two-dimensional visualizations in which the distance between points reflects their similarity in 400-dimensional space (37).

Modeling

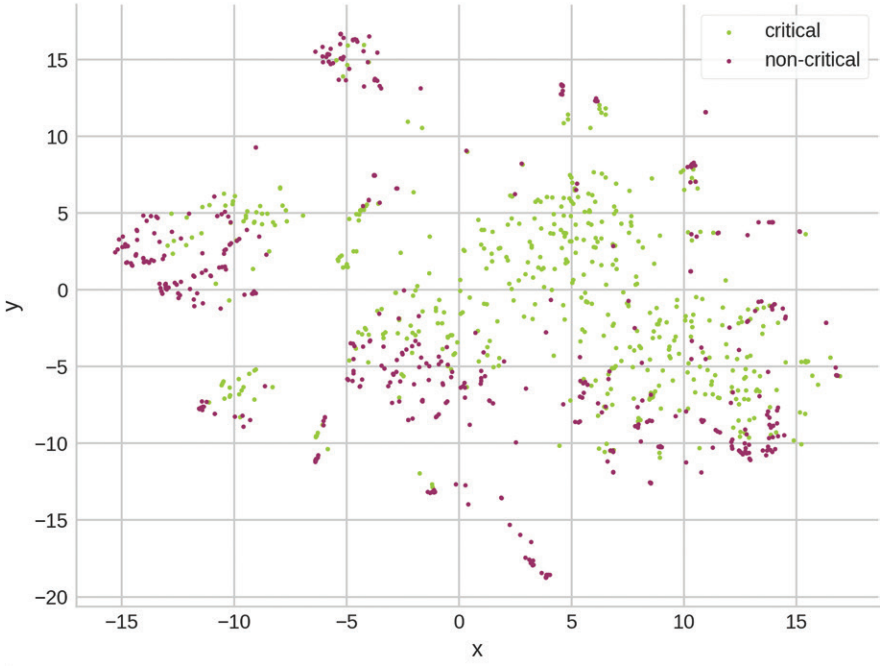
We attempted to predict human-generated reference-standard document labels using the features described above for the 1004 labeled reports of the head CT corpus. Features were entered into a Lasso logistic regression to predict labels that occurred at least 20 times. All statistical analyses were performed with scikit-learn 0.18.1 (38). Lasso logistic regression differs from standard logistic regression in that it introduces a penalty for assigning weight to regression coefficients (39). This is practically useful, as such models can take a very large number of features as input and assign regression coefficients of 0 to many uninformative features, effectively ignoring them. Nonzero regression coefficients are assigned only to those variables that contribute sufficient accuracy to the model's predictions. Sixty percent (602 of 1004) of the labeled reports were used to train regression models, and 40% (402 of 1004) of the labeled reports were held out to validate the accuracy of the models. Final models were scored by area under the receiver operating characteristic curve (AUC), and each model's mean, minimum, maximum, and standard deviation of AUC was reported for *(a)* all labels in aggregate, *(b)* critical labels only, and *(c)* a binary variable indicating whether a report contained a critical finding. Sensitivity, specificity, accuracy, and F1 score—defined as (two times precision times recall)/(precision plus recall), where precision refers to positive predictive value and recall refers to sensitivity—were reported for the best-performing model's *(a)* predictions of all labels and *(b)* identification of reports containing critical findings. We illustrate held-out AUC curves for selected labels. To illustrate the mechanics of unigram BOW modeling, the average, minimum, and maximum number of nonzero terms used by the unigram BOW model across all labels were calculated, and examples of terms learned for specific labels were reported.

Code to enable readers to replicate these methods on their own data will be made available for download at *https://github.com/aisinai*.

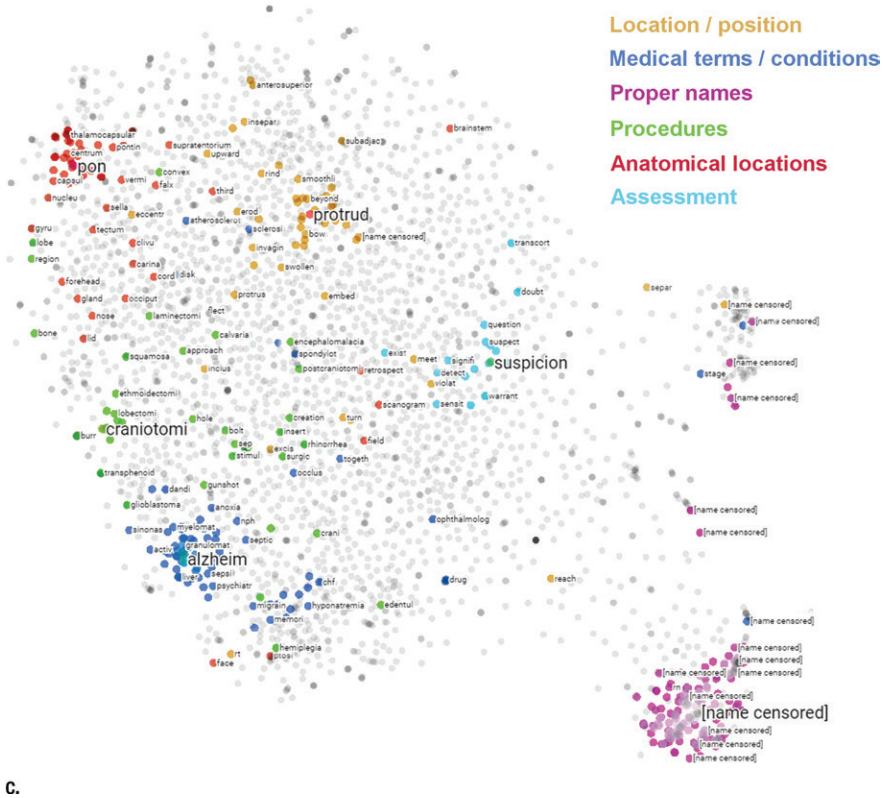
Figure 3



a.



b.



c.

Figure 3: **(a)** The unsupervised embedding models on the head CT corpus learn vectors corresponding to each word; remarkably, when these vectors are added algebraically, they reflect the underlying semantics of these concepts. For more information, see reference 50. **(b)** Utilizing *t*-SNE to visualize the embedding space and our reference-standard labels, we can observe a centralization of critical head documents based on their average word-embedding vectors, indicating that documents with critical findings tend to be clustered together in this space. **(c)** Also utilizing *t*-SNE in an unsupervised fashion, we can note that the head CT corpus has well-defined clusters of shared semantic meaning. *t*-SNE = *t*-distributed stochastic neighbor embedding.

Results

Comparative Lexical Analysis

The English language corpora had more unique words, more unique bigrams, and higher TTR scores relative to all three radiology corpora. Notably, the average number of words in the head CT corpus was 5997 words across 96303 reports, compared with 26546 across 11887 paragraphs of Reuters, with respective TTR scores of 0.011 and 0.046. Similarly, the English language corpora had an average Yule I of 103 and 76 for the Reuters and Gutenberg corpora, respectively, whereas for each radiology corpus, the Yule I was lower than 40. On the normalized relative complexity measure, the radiology corpora were 10 to 30 times less complex than Reuters (Table 1). Despite accounting for only 1304 summaries, the I2B2 UTHEALTH had an average of 32014 unique words, a Yule I of 271, and an average normalized relative complexity of 1.36 times that of the Reuters corpus.

NLP Modeling

Most labels occurred infrequently (median incidence, 5.4% [54 of 1004]). Excepting the most common label (presence of any abnormality, 81.6% [819 of 1004]), the next most common label (sinus disease) was present in 29.4% (295 of 1004) of reports.

Table 2 presents the results of the best-performing feature extraction and model pairs. Notably, a simple BOW approach with unigrams with a Lasso logistic regression achieved an average AUC of 0.950 ± 0.046 (standard deviation) on all labels and 0.951 ± 0.040 on critical labels. Utilizing the average word embedding vectors in a Lasso logistic regression yielded an AUC of 0.917 ± 0.056 on all labels and 0.935 ± 0.055 on critical labels. The best-performing LDA model utilizing 200 topics and unigrams achieved an AUC of 0.872 ± 0.091 on all labels and 0.914 ± 0.060 on critical labels. The best-performing overall model (a combination of BOW and average word embedding vector) achieved an AUC of 0.957 ± 0.039 on all labels, 0.959 on critical labels, and 0.966 in identifying reports containing any critical label (Fig 2). It achieved an average sensitivity and specificity for all labels of 90.25% (1898 of 2103) and 91.72% (18351 of 20007), respectively, and identified reports containing critical findings with a sensitivity and specificity of 92.59% (175 of 189) and 89.67% (191 of 213), respectively (Table 3).

In the unigram BOW model, out of 3291 possible unigram tokens, our regression model utilized an average 35.8 tokens (range, six to 86 tokens) by assigning a nonzero weight to their regression coefficients. Certain models, such as that for "CNS edema," used a limited number of tokens (n = 18) and tended to put disproportionate weight on a single token whose appearance was strongly associated with the finding (regression coefficient of 4.96 for "edema," next most influential coefficient, −0.84 "acut"). Other models, such as that for "stroke/infarction/ischemia," used more tokens (n = 65) and assigned substantial regression coefficients to multiple tokens (five most influential tokens: 3.65 for "lacunar," 2.48 for "infarct," 1.46 for "stroke," 1.34 for "chronic," and 1.19 for "hypodense"). Unigram BOW models for "edema" and "stroke/infarction/ischemia" achieved AUCs of 0.991 and 0.954, respectively.

In addition to their accuracy as classifiers, the embedding models and topic models provided a means of interpreting the underlying semantic structure of the radiology corpora. By looking at the similarity between average word embedding vectors, simple clinical and semantic relationships can be recapitulated (Fig 3a). By projecting the reports into the embedding space utilizing the average word embedding vector model and then visualizing it with t-SNE, we can note an underlying structure of critical CT findings being located centrally, with noncritical findings on the periphery (Fig 3b). Similarly, visualization of the embedding space itself for the head CT corpus allows us to note distinct, semantically-related groupings of words corresponding to location, medical terms and conditions, proper names, procedures, anatomic locations, and radiologist assessment (Fig 3c). Topic models proved to be one of the least accurate means of generating document classifiers, with a best average AUC of 0.872. However, topic models did have the advantage of being highly interpretable, with the ability to specifically survey the individual words that constitute each topic (Fig 4).

Discussion

The annotation of large radiology report corpora to facilitate large-scale research in radiology with machine learning and deep learning is itself a nontrivial problem in NLP. Traditional rule-based approaches can achieve impressive results but may be hard to generalize outside of the training sets on which the rules are constructed (3,6,11). We found that simple unigram BOW features in a Lasso logistic regression performed well at labeling clinical entities in a large radiology corpus. Furthermore, distinct from rule-based methods, which are typically hand engineered for a particular type of report, machine learning approaches such as those used here can be easily applied to a wide variety of radiology reports.

Our accuracy for critical findings of 91.04% demonstrates a substantial improvement on a past effort in 2000 to automatically code head CT head reports that achieved 84% accuracy (40). A recent study applied machine learning classification (support vector machine with BOW) to knee MR imaging reports and achieved a same-hospital F1 score of 0.903–0.921 in identifying a composite variable indicating findings of interest across 2454 reports (41). Our results demonstrate that it is possible to achieve a similarly high F1 score (0.907) in identifying reports containing critical findings in a completely different domain of radiology.

Our complexity analysis results suggest that part of the success of simple models in the radiology context is specific to the actual lexical and semantic structure of the language of radiology reporting. Despite their relative sizes being almost an order of magnitude smaller than the radiology corpora, the English language corpora had more unique words, more unique

Figure 4

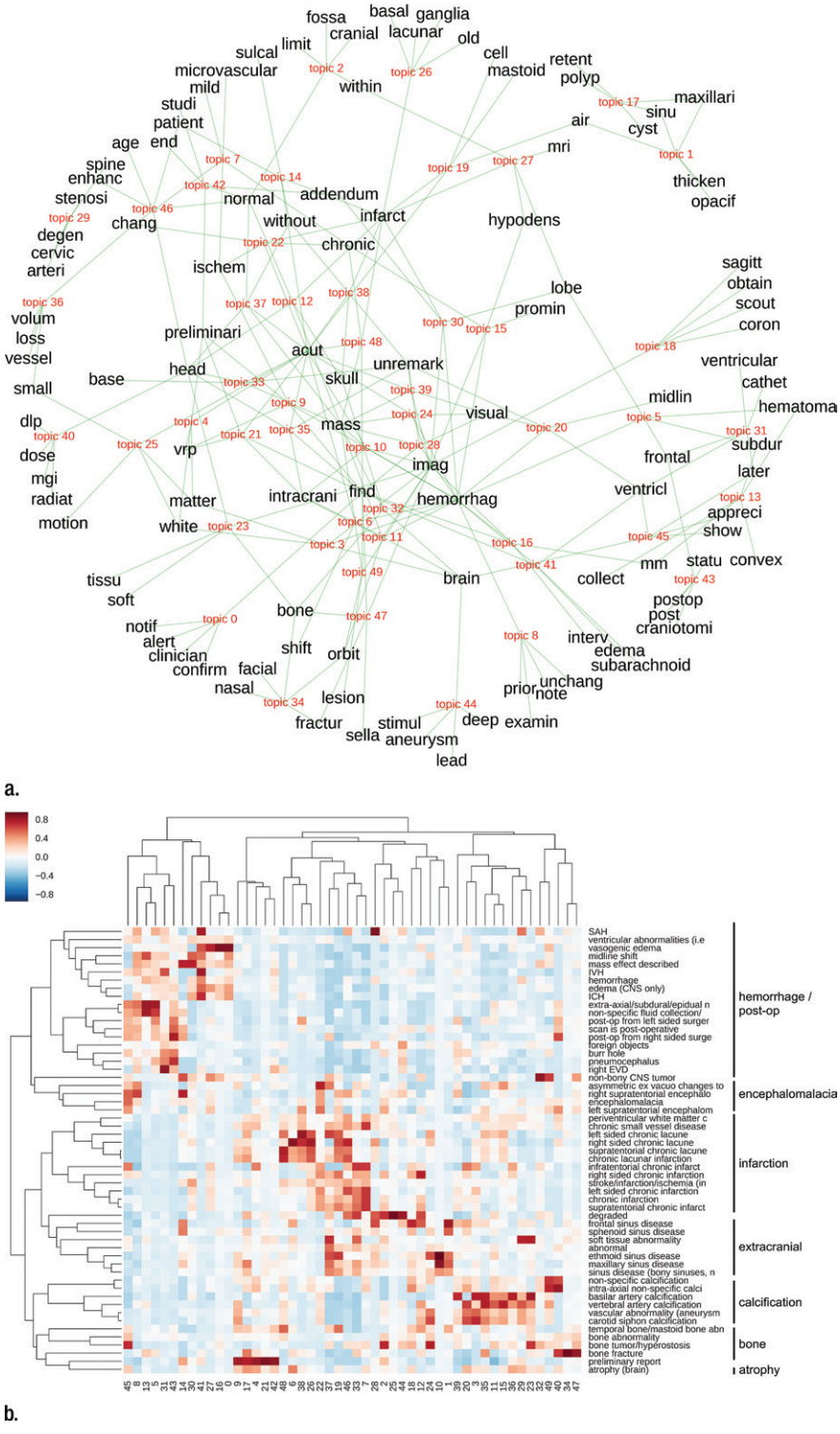


Figure 4: The Latent Dirichlet allocation topic models provide a highly transparent means of document classification. **(a)** Each topic is illustrated in red text, and lines are drawn to each of the five words most commonly associated with that topic. Certain common words were shared by multiple topics (eg, "acute"), and these words appear more centrally on the diagram. **(b)** Each box represents the average relative weight of a topic (bottom axis) in reports that were positive for a given label (right axis); red = relatively high weight, and blue = relatively low weight. Labels are sorted vertically so that labels with more similar topic weight distributions appear in proximity to each other. When used to assign average topic weights to reports, topics can be readily demonstrated to have captured the underlying clinical semantics both broadly and in specific cases.

bigrams, and higher TTR scores relative to all three radiology corpora. Similarly, the I2B2 UTHEALTH corpus was easily the most complicated corpus, which is suggestive of a substantially greater difficulty for performing NLP in the general medical context relative to radiology or even standard English language cases. The limited anatomy and pathology described in any given type of report, as well as the use of automatically generated text and dictation software, are likely both major contributors to the lexical simplicity of radiology. Furthermore, the highly specialized nature of each imaging modality lends itself to the de facto creation of sublanguages each with a relatively low degree of lexical and semantic complexity. This is likely to become even truer going forward with the wider adoption of standardized reporting protocols (eg, LI-RADS, PI-RADS) (42).

The finding that relatively simple BOW approaches performed well invites the question of whether more sophisticated and computationally expensive approaches are truly required for this application. One strength of these more sophisticated models is that they offer additional characterization of the reports that goes beyond their value as predictive features, and human interpreters can often discover interesting patterns in analyzing these

results (19,34). One of the more unexpected findings, for example, was the distinct groupings associated with physician names, procedures, and anatomic locations. Some recent studies have reported utilizing this finding for performing de-identification, where a similar analysis is performed to identify regions of patient names or other protected health information, and then subsequently including those words as stop words or excluding those regions (43,44).

While we report aggregate AUC numbers for our classifiers, we note that there was substantial variation in the ability of our approach to recover different labels. This reflects the fact that certain labels (eg, "edema") are referred to with very consistent language in most reports, whereas other labels (eg, "stroke/infarction/ischemia") may be referred to with a wider variety of language. With an increasing trend toward standardization of reporting language, we anticipate that more labels will become reliably automatically recoverable in the future.

The method described in this article can be used to generate a large set of automatically labeled radiology reports based on a small set of manually labeled reports. Because these labels are based purely on the text contained in the reports and do not incorporate any machine-derived features learned from the corresponding imaging, they can be used as an independent set of labels to train deep learning–based image classification models. Work in other domains has shown that using large data sets with inferred labels for image classification tasks can deliver excellent results, as the larger amount of training data more than compensates for inaccuracy introduced into the inferred training labels (45–49). Additionally, on generating a set of labels, active learning or semisupervised learning approaches can be utilized as a further means of de-"noising" the data set prior to subsequent supervised learning (49).

An important limitation to our work was that all reports included in the radiology report corpus were derived from only two hospitals. The exact classifiers learned in this work may not generalize to external data sets. However, provided that a subset of the reports in a new corpus can be manually labeled, this same method can be implemented to infer corpus-specific labels for reports in any radiologic corpus.

In conclusion, because of the highly structured language of radiology, straightforward machine learning–based approaches can achieve state-of-the-art classification results in text corpora of radiology reports, generating large data sets of labeled reports needed for applications such as deep learning.

Acknowledgments: We thank Burton Drayer, MD, and Errol Gordon, MD, for their contribution to and support of this work. Deidentified clinical records used in this research were provided by the I2B2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Ozlem Uzuner, PhD, I2B2, and SUNY.

Disclosures of Conflicts of Interest: J.Z. disclosed no relevant relationships. **M.P.** disclosed no relevant relationships. **J.T.** disclosed no relevant relationships. **M.B.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is employed as a fellow at Verily Life Sciences. Other relationships: disclosed no relevant relationships. **J.S.** disclosed no relevant relationships. **A.S.** disclosed no relevant relationships. **A.C.** disclosed no relevant relationships. **J.B.** disclosed no relevant relationships. **J.L.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: was an employee of Verily Life Sciences at the time of this work; is an adviser/consultant for Google; is currently an employee of Merck. Other relationships: disclosed no relevant relationships. **E.K.O.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is employed as a postdoctoral fellow at Verily Life Sciences. Other relationships: disclosed no relevant relationships.

References

1. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015;350:h1885.
2. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;306(8):848–855.
3. Jung K, LePendu P, Iyer S, Bauer-Mehren A, Percha B, Shah NH. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. J Am Med Inform Assoc 2015;22(1):121–131.
4. Huang SH, LePendu P, Iyer SV, Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. J Am Med Inform Assoc 2014;21(6):1069–1075.
5. Agarwal V, Podchiyska T, Banda JM, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc 2016;23(6):1166–1173.
6. Carrell DS, Halgrim S, Tran DT, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol 2014;179(6):749–758.
7. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6(1):30.
8. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016;279(2):329–343.
9. Dreyer KJ, Kalra MK, Maher MM, et al. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 2005;234(2):323–329.
10. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7(6):593–604.
11. Percha B, Nassif H, Lipson J, Burnside E, Rubin D. Automatic classification of mammography reports by BI-RADS breast tissue composition class. J Am Med Inform Assoc 2012;19(5):913–916.
12. Solti I, Cooke CR, Xia F, Wurfel MM. Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches. Proceedings IEEE Int Conf Bioinformatics Biomed 2009;2009:314–319.
13. Mitchell TM. The discipline of machine learning, Vol 3. Pittsburgh, Pa: Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.
14. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.

15. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. arXiv [cs.CV]. 2014.

16. Shin HC, Lu L, Kim L, Seff A, Yao J, Summers RM. Interleaved text/image Deep Mining on a large-scale radiology database. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.

17. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annu Symp Proc 2006:269–273.

18. Mcauliffe JD, Blei DM. Supervised topic models. In: Platt JC, Koller D, Singer Y, Roweis ST, eds. Advances in neural information processing systems 20. Red Hook, NY: Curran Associates, 2008; 121–128.

19. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. arXiv [cs.CL]. 2013.

20. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of word representations in vector space. arXiv [cs.CL]. 2013.

21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377–381.

22. Stubbs A, Uzuner Ö. Annotating risk factors for heart disease in clinical narratives for diabetic patients. J Biomed Inform 2015;58(Suppl):S78–S91.

23. Lewis DD. Reuters-21578 text categorization test collection, distribution 1.0 http://www.daviddlewis.com/resources/testcollections/reuters21578/. Published 1997. Accessed April 13, 2017.

24. Project Gutenberg. http://www.gutenberg.org. Published 2016. Accessed April 13, 2017.

25. Stubbs A, Kotfila C, Xu H, Uzuner Ö. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2. J Biomed Inform 2015;58(Suppl):S67–S77.

26. McAuley J, Leskovec J. Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of the 7th ACM conference on Recommender systems. ACM, 2013; 165–172.

27. Yule GU. On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. Biometrika 1939;30(3/4):363–390.

28. Montemurro MA, Zanette DH. Universal entropy of word ordering across linguistic families. PLoS One 2011;6(5):e19875.

29. Porter MF. An algorithm for suffix stripping. Program. 1980;14(3):130–137.

30. Ghahramani Z. Unsupervised learning. In: Bousquet O, von Luxburg U, Rätsch G, eds. Advanced lectures on machine learning. Berlin, Germany: Springer-Verlag, 2004.

31. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3(Jan):993–1022.

32. Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA, 2010; 45–50.

33. Bishop CM. Pattern recognition and machine learning. New York, NY: Springer-Verlag, 2006.

34. Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM. Reading tea leaves: how humans interpret topic models. In: Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. Advances in Neural Information Processing Systems 22. Red Hook, NY: Curran Associates, 2009; 288–296.

35. DARIAH-DE. NLP Based Analysis of Literary Texts. https://github.com/stefanpernes/dariah-nlp-tutorial. Accessed April 13, 2017.

36. Viertel VG, Trotter SA, Babiarz LS, et al. Reporting of critical findings in neuroradiology. AJR Am J Roentgenol 2013;200(5):1132–1137.

37. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(Nov):2579–2605.

38. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(Oct):2825–2830.

39. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Stat Methodol 1996;58(1):267–288.

40. Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res 2000;33(1):1–10.

41. 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 2017;208(4):750–753.

42. Castellanos SH, Gonzalez-Aguirre A, Ibargüengoitia MC, Vazquez S, Lamadrid JV. BI-RADS, C-RADS, GI-RADS, LI-RADS, Lu-RADS, TI-RADS, PI-RADS. The long and winding road of standardization. Proceedings of the European Congress of Radiology. 2014.

43. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017;24(3):596–606.

44. Kassaie B. De-identification In practice. arXiv [cs.CL]. 2017.

45. Joulin A, van der Maaten L, Jabri A, Vasilache N. Learning visual features from large weakly supervised data. arXiv [cs.CV]. 2015.

46. Krasin I, Duerig T, Alldrin N, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. https://github.com/openimages. Accessed March 22, 2017.

47. Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, et al. YFCC100M: The new data in multimedia research. arXiv [cs.MM]. 2015.

48. Bai Y, Yang K, Yu W, Xu C, Ma WY, Zhao T. Automatic image dataset construction from click-through logs using deep neural network. In: Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015; 441–450.

49. Collins B, Deng J, Li K, Fei-Fei L. Towards scalable dataset construction: an active learning approach. In: Computer Vision – ECCV 2008. Berlin, Germany: Springer, 2008; 86–98.

50. Mikolov T, Yih WT, Zweig G. Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Ga: Association for Computational Linguistics, 2013; 746–751.