Automatic Classification of Critical Findings in Radiology Reports

Aditya Tiwari[1], Samah Fodeh[2], Steven Baccei[3], Max Rosen[3]

[1]University of Massachusetts Amherst, [2]Yale University, [3]University of Massachusetts Medical School

***Abstract*— Communication of "actionable" findings in radiology reports is an important part of high quality medical care. Distinguishing radiology reports with "actionable" findings from other reports is currently a function of the radiologist and largely a manual process. This paper describes a system for automatic classification of patient's radiology reports as it relates to the degree of severity of "actionable" findings provided by the radiology department at University of Massachusetts Medical School. This is done by using machine learning classifier on text based features. Several machine learning classification algorithms are evaluated and compared. Random forest classifier performed the best in this case while other classification methods also performed decently.**

***Index Terms*— Critical Findings, Machine Learning, Multi-Class Classification, Radiology, Random Forest**

I. INTRODUCTION

Accurate classification and communication of "actionable" findings on diagnostic radiology reports is essential for relaying critical patient information to referring physicians. Cases with actionable findings are urgent, requiring prompt, or in some instances immediate attention by the referring clinician. The severity of a patient's condition can be inferred using their radiology report(s) which contain the detailed information about the patient's imaging findings and diagnosis.

Patients' radiology reports are text records of the Radiologist's interpretation of the radiology examination, which contain important information including information about the severity of the imaging findings. These reports also include information such as the patient's consultation history, examination details, indication, technique, findings, impression etc. and provide an assessment about the patient's health status which then can be used by referring clinicians to direct and organize further care for the patient based on the urgency of the findings. Since there are a large volume and variety of medical records, including radiology reports, it can be challenging and cumbersome to manually classify reports. Thus, automatic classification of such records can save an enormous amount of time and potentially improve patient care by delivering "actionable" findings to a patient's clinician both faster and in a "highlighted" fashion.

To classify radiology reports using their text contents, we have used machine learning algorithms. As per the wiki definition, machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed. These algorithms can learn from data and then can be used for predictions. Machine Learning has become part of healthcare through many applications such as breast tumor classification [11], extraction of clinical entities from hospital discharge entities [8], and detection of suspicious access of electronic health records (EHRs) [1]. In particular, it has also been widely applied in radiology report mining. Berry et.al [3] has compared various Information Retrieval (IR) and machine learning based classifiers for categorization of wrist x-ray reports. The dataset had 751 textual reports which don't indicate about the generalizability of the classifier on bigger dataset. Similarly, in [4] identification of limb features from radiology reports is done using Naive Bayes and SVM classifiers by training classifiers on just 99 reports. Bijoy et.al [10] has tried to categorize radiology reports to identify fractures. This work is based on search of specific words like "fracture" in the text report and is thus specific to that domain. In a similar study by Nguyen's [9], characteristics of cancer were extracted based on rule based classifier which uses cancer related nomenclature. In [13], two machine learning techniques, i.e. Naive Bayes and dynamic language model (DLM), were utilized to classify radiology reports to facilitate identification of radiological examinations for retrospective research projects.

In this paper, we propose classifying radiology reports based on the severity of findings. We define here the levels of finding severity: and our method utilizes features derived from radiology reports. The features used in the classifier are not limited to any specific disease. On the contrary, we utilize all contents in the radiology reports to classify the level of urgency of the report's findings. Our classifier outperformed various classification algorithms, as we show in the Results Section.

The remainder of the paper is organized as follows. Section II discusses the methodology adopted for classification. Section II.B provides the details for feature extraction from the reports. These features are then used for classification. Details about the classification techniques are given in section II.C Section III discuss the results by various classification techniques and their comparison. Section IV then summarizes the conclusion in the paper.

II. METHODOLOGY

Radiology reports are dictated by radiologists in text format. In our approach, we extract text features and use them as the basis for classification. The text, thereby, needs to be preprocessed before proceeding with the classification task. For classification, the dataset is divided into training and testing sets, and the classification model is built using the training set and evaluated with the test set.

A. Data

The dataset is provided by the Radiology Department at University of Massachusetts Medical School. It contains a sample of patient's

This work has been funded by the Radiology department at University of Massachusetts Medical School.

radiology reports from 1/1/2010-1/1/2016. These reports are classified into four categories according to urgency, i.e. Normal, Yellow, Orange and Red. Normal Category means that the patient's findings do not require immediate attention. Similarly, Red category belongs to the patients, whose Radiology reports contain findings that necessitate urgent attention. We have received from the Radiology department two datasets:

1. *Binary dataset*: consists of two labels; Normal and Red. In this dataset, within the normal category, there were 939 reports and for Red category there are 953 reports. After document preprocessing, we identified 1004 features on which classification is performed.
2. *Multi-labeled dataset*: consists of four labels: Normal, Yellow, Orange and Red. The numbers of reports for each of the categories are 939, 641, 591 and 953 respectively. The total number of features is 1004, after document preprocessing. Yellow was defined as *"findings that are typically unexpected but generally do not require any immediate treatment or other action but in the long term could be very significant"* and Orange was defined as *"clinically significant findings that generally explain a patient's acute presentation and require specific medical or surgical treatment but are not imminently life threatening"*.

B. Feature Engineering

The radiology reports were processed to generate text features which can then be used for training the classifier. We formed a representation for the reports using "bag of words" (BOW) model. In BOW, each text report is represented as a vector of words included in that report. It accounts for frequency of words in the text report and ignores the ordering of words and grammar in the text. We designed a pipeline of the following operations on the text:

- Tokenization: separate/tokenize words in a report using spaces as delimiters.
- Stop words removal: removed the common words in English using a default stop list of English words.
- Stemming: We then transformed words into their roots using a python implementation of porter stemmer.

The processed result of the pipeline is then used to generate text features. We experimented with 2 types of feature weights: term frequency and *TF-IDF* (term frequency- inverse document frequency) weights. In term frequency, we combine vectors of all reports into a matrix. Words in the matrix are weighted based on their frequencies. An entry $c_{i,j}$ in the matrix is the frequency of having the word *i* in report *j*. In *TF-IDF*, word frequencies within reports are weighted down by frequency of words in all reports in the dataset (inverse document frequency). The formula for *TF-IDF* for a term t in a document is defined as:

$$Tf - Idf = tf(t) * idf(t)$$

Where the first term *tf(t)* is frequency of the term in the given document. And the second term, *idf(d, t),* is defined as: *idf(d, t)* $=\log\left[\frac{(1+n)}{1+df(d,t)}\right]+1$, where n is the document frequency and *df(d,t)* is the number of documents that contain term t.

In our initial experiments, term frequency features performed better than *TF-IDF* features. The experiment is mentioned in Results section. The *TF-IDF* features are in general beneficial to reduce the weights of stop words, which are frequently appearing words in most documents. In our pipeline we have already removed stop words, thus *TF-IDF* are not very helpful. Also, *TF-IDF* features are more prominent in the recommender system where ranking of documents is done based on feature weights. In classification problems, *TF-IDF* might reduce weights of features which are significant to a class in the training dataset. Thus our remaining experiments are based on term frequency features.

C. Classification

In a supervised classification problem, a machine learning model is developed using training examples for which the labels/classes are already known. The classification model learns the attributes and features that separate the data points into the different classes. Once developed, the classifier can label new and previously unseen data points with the appropriate class. We have tried various classification algorithms including: Decision Trees [6], Random Forest [7], Support Vector Machine (SVM) [2], Linear Discriminant Analysis (LDA) [12], Naive Bayes [5], K-Nearest Neighbor (KNN) [5] . These classifiers were learned using the training dataset and their prediction accuracy was measured on the test dataset. Since, we have different datasets one binary and the other with multiple classes we will employ two classification approaches:

- **Binary Classification:** This approach classifies data into two classes. The classification model, using the labeled data points, learns the two classes and the essential data attributes for data separation. in our binary dataset we have two classes: Normal and Red.
- **Multiclass classification**. This class of problem classifies data into multiple classes (more than 2). In this multiclass classification problem, we assign each data point a single label out of four possible options. To solve a multiclass classification problem, we transform it into sub problems of binary classification. That is a class is solved against all other classes in a binary classification problem. This task is repeated for each class. In one of the dataset, we have 4 classes i.e. Normal, Yellow, Orange, Red.

III. RESULTS

We divided the data into training and test sets. The data was divided uniformly from all the categories. Training set consists of 70 % entries and rest is considered as the test set for each class. The training set is used for training and validation of classifiers.

A. Cross Validation

We have adopted **10 fold cross validation** for model selection. In this we randomly divide the training dataset in 10 equal parts. 9 of the 10 parts are used for training the model and remaining one part is used for validation of the trained model. This process is repeated 10 times with each of the parts used exactly once for validation. We adopted a grid search for various sets of hyper parameters for each model and adopted a model which gives us the best average training accuracy. The test set is used for reporting results.

B. Evaluation

The classification performance of the classifiers is evaluated using standard metrics including Overall Accuracy, Precision, Recall and F-Measure. They are defined as follows:

$$Overall\ Accuracy = \frac{\text{Correctly classified instances}}{\text{Total test instances}} \quad (1)$$

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F - Measure = \frac{2*Precision*recall}{Precision+Recall} \quad (4)$$

In the above stated equations, TP, FP and FN correspond to number of true positives, false positives and false negatives respectively. It means that for a given class *c*, TP denotes the number of reports which are predicted as class *c*, and according to test data they are annotated as class *c* as well; FP denotes the number of reports that were labeled as *c* by the classifier but this label does not match their annotated class in the test set; FN denotes the number of reports that have class *c* according to the test set but are labeled as some other class by the classifier. These standard metrics are used for reporting the performance of binary and multiclass-classifiers.

C. Experiments & Results

C.1 Binary Classification: Normal and Red

Following tables (1-7) showcase the results for the binary class classification performance for various classification algorithms. Here the two classes are, Normal and Red.

Table 1: Random Forest, Overall Accuracy: 98.24 %

Class	Precision	Recall	F-measure
Normal	0.97	0.99	0.98
Red	0.99	0.97	0.98

Table 2: Decision Trees, Overall Accuracy: 95.77 %

Class	Precision	Recall	F-measure
Normal	0.94	0.98	0.96
Red	0.98	0.94	0.96

Table 3: SVM, Overall Accuracy: 96.65 %

Class	Precision	Recall	F-measure
Normal	0.96	0.98	0.97
Red	0.98	0.95	0.97

Table 4: Linear Discriminant Analysis (LDA), Overall Accuracy: 84.51 %

Class	Precision	Recall	F-measure
Normal	0.84	0.85	0.84
Red	0.85	0.84	0.85

Table 5: Naive Bayes, Overall Accuracy: 94.19 %

Class	Precision	Recall	F-measure
Normal	0.91	0.98	0.94
Red	0.97	0.91	0.94

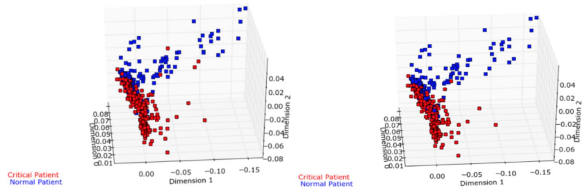
Table 6: K-Nearest Neighbor (KNN), Overall Accuracy: 85 %

Class	Precision	Recall	F-measure
Normal	0.78	0.96	0.86
Red	0.95	0.74	0.83

Table 7: AdaBoost, Overall Accuracy: 97.01 %

Class	Precision	Recall	F-measure
Normal	0.95	0.99	0.97
Red	0.99	0.95	0.97

Fig. 1 is a visualization of the classification performance of random forest algorithm for two classes. This figure is generated by reducing the dimensions of the text features to 3 using Principle Component Analysis (PCA); a well-known dimensionality reduction approach. Then all the instances in the test data set are plotted on the 3 dimensional spaces. Fig. 1(a) contains the visualization of test instances with actual classes that are known apriori. In this normal category instances are marked in blue and red category instances are marked in red color. Fig. 1(b) contains the visualization of test instances indicating predicted classes. Instances which are predicted as normal category are marked in blue and the ones which are predicted as red category are marked in red color.



(a) Actual classes (b) Predicted Classes

Fig. 1: Visualization of random forest classifier's performance for two classes

C.2 Multiclass Classification: Normal, Yellow, Orange and Red

Following tables (8-14) explain the results of the multi-class classification where the classes are, Normal, Yellow, Orange and Red. The results are evaluated for various classification algorithms.

Table 8: Random Forest, Overall Accuracy: 94.36 %

Class	Precision	Recall	F-measure
Normal	0.95	0.99	0.97
Yellow	0.96	0.89	0.92
Orange	0.92	0.91	0.91
Red	0.95	0.95	0.95

Table 9: Decision Trees, Overall Accuracy: 93.61 %

Class	Precision	Recall	F-measure
Normal	0.94	0.99	0.97
Yellow	0.95	0.89	0.92
Orange	0.87	0.92	0.89
Red	0.96	0.93	0.94

Table 10: SVM, Overall Accuracy: 92.33 %

Class	Precision	Recall	F-measure
Normal	0.94	0.98	0.96
Yellow	0.90	0.92	0.91
Orange	0.90	0.85	0.87
Red	0.94	0.91	0.92

The results for multiclass classification problem also point that random forest outperformed other classifiers. Worst results are

achieved by KNN classifier.

Table 11: Linear Discriminant Analysis (LDA), Overall Accuracy: 89.35 %

Class	Precision	Recall	F-measure
Normal	0.86	0.94	0.90
Yellow	0.92	0.88	0.90
Orange	0.89	0.81	0.85
Red	0.91	0.91	0.91

Table 12: Naive Bayes, Overall Accuracy: 79.66 %

Class	Precision	Recall	F-measure
Normal	0.92	0.89	0.90
Yellow	0.73	0.75	0.74
Orange	0.60	0.83	0.70
Red	0.92	0.72	0.80

Table 13: K-Nearest Neighbor (KNN), Overall Accuracy: 73.80 %

Class	Precision	Recall	F-measure
Normal	0.67	0.95	0.79
Yellow	0.77	0.64	0.70
Orange	0.67	0.60	0.63
Red	0.88	0.69	0.77

Table 14: AdaBoost, Overall Accuracy: 93.18 %

Class	Precision	Recall	F-measure
Normal	0.91	0.99	0.95
Yellow	0.93	0.92	0.92
Orange	0.95	0.87	0.91
Red	0.95	0.92	0.93

Previously in section II.B, we mentioned that results from term frequency features were better than tf-idf in initial experiments. We experimented with random forest with tf-idf features for multiclass classification. Results are presented in table 15 and results with term frequency features (table 8) are decently better than with tf-idf features.

Table 15: Random Forest **(TF-IDF Features)**, Overall Accuracy: 86.47 %

Class	Precision	Recall	F-measure
Normal	0.91	0.99	0.95
Yellow	0.79	0.81	0.80
Orange	0.77	0.67	0.71
Red	0.92	0.90	0.91

IV. Discussion

We have presented a system for automatic classification of patient's radiology reports, which are provided by the radiology department at University of Massachusetts Memorial Medical Center. A critical finding on a patient's radiology report can be given appropriate importance if the critical information can be mined precisely from the diagnostic reports in timely manner. This paper presents an approach to extract patient critical findings from the patient's diagnostic report by using machine learning techniques. The results achieved on the given dataset are impressive and thus can be used in practical scenarios to classify reports.

The diagnostic reports are text documents that contain information like patient's consultation history, examination details, findings etc. These parameters often contain important details related to each patient's health. These text reports are processed to get text based features which are then used with various machine learning algorithms. Machine learning algorithms have previously been used in healthcare for solving various problems like breast tumor classification, identification of limb features from radiology reports, identification of fractures and many more. We have experimented with a range of binary and multiclass classification algorithms. This study shows that, on the given diagnostic reports dataset, random forest classifier performs better than the other classifiers for text classification in both, binary and multi-class, cases. A random forest is a classifier that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. Also, it is fast and scalable. Although the results achieved are effective enough to be application ready, there are various recent concepts in machine learning domain that can also be explored. For example, deep learning is the current area in machine learning, which is most active. It has been used to solve many difficult problems in machine learning like computer vision, natural language processing, robotics etc. Such advanced techniques can be experimented in future studies.

References

[1] Aziz A Boxwala, Jihoon Kim, Janice M Grillo, and Lucila Ohno-Machado. Using statistical and machine learning to help institutions detect suspicious access to electronic health records. Journal of the American Medical Informatics Association, 18(4):498{505, 2011.

[2] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, pages 144{152. ACM, 1992.

[3] De Bruijn B , Cranney A , O'Donnell S , Martin JD , Forster AJ . Identifying wrist fracture patients with high accuracy by automatic categorization of x-ray reports . JAMIA . 2006 ; 13 (6): 696 – 698 .

[4] Guido Zuccon, Amol S. Wagholikar, Anthony N. Nguyen, Luke Butt, Kevin Chu, Shane Martin, and Jaimi Greenslade. Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology. In AMIA Summits on Translational Science 2013, pages 300{304, San Francisco, California, 2013. American Medical Informatics Association.

[5] Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer, 2017. Print.

[6] J. Ross Quinlan. Induction of decision trees. Machine learning, 1(1):81{106, 1986.

[7] Leo Breiman. Random forests. Machine learning, 45(1):5{32, 2001.

[8] Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. Journal of the

American Medical Informatics Association, 18(5):601{606, 2011.

[9] Nguyen A , Moore J , Lawley M , Hansen D , Colquist S . Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications . Health Informatics Conference . 2011 : 117 – 124 .

[10] Thomas BJ , Ouellette H , Halpern EF , Rosenthal DI . Automated computer-assisted categorization of radiology reports . American Journal of Roentgenology . 2005 ; 184 (2): 687 – 690 .

[11] Tim W. Nattkemper, Bert Arnrich, Oliver Lichte, Wiebke Timm, Andreas Degenhard, Linda Pointon, Carmel Hayes, and Martin O. Leach. Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods. Artificial Intelligence in Medicine, 34(2):129 { 139, 2005.

[12] William R Klecka. Discriminant analysis. Number 19. Sage, 1980.

[13] Zhou Y1, Amundson PK, Yu F, Kessler MM, Benzinger TL, Wippold FJ, Automated classification of radiology reports to facilitate retrospective study in radiology. J Digit Imaging. 2014 Dec;27(6):730-6. doi: 10.1007/s10278-014-9708-x.