COMPUTATIONAL METHODS FOR CATEGORIZING UNSTRUCTURED DATA RELATED TO PEDIATRIC APPENDICITIS WITHIN ELECTRONIC MEDICAL RECORDS

by

BRITTANY NORMAN

(Under the Direction of Khaled Rasheed)

ABSTRACT

Analyzing free-form medical data such as pathology reports or physicians' notes and comments presents additional challenges. Unlike with structured data (e.g., numerical, check-boxes, ICD-10 codes), there are countless ways that physicians could express the same concept in unstructured text. In this thesis, computational techniques were explored for automating the categorization of medical documents related to pediatric appendicitis. In the first project, a computational model was built to detect emergency department notes which contained features of Pediatric Appendicitis Score. This model achieved a 0.8391 F-Score compared to human performance and also outperformed the previous computational method (0.3435 F-Score). In the second project, a model was constructed to identify appendectomy pathology reports which were negative for appendicitis. This model obtained an F-Score of 0.9960. In many cases, hospitals rely on manual chart review for such tasks; this thesis presents an alternative computational approach using statistical natural language processing.

INDEX WORDS: Appendectomy, Appendicitis, Artificial Intelligence, Big Data, Biomedical Informatics, Business Intelligence, Chart Review, Classification, Computer Science, Data Mining, Data Science, Document Classification, Electronic Health Record, Electronic Medical Record, Emergency Medicine, Health Informatics, Health Sciences, Health Technology, Information Retrieval, Logistic Regression, Machine Learning, Medical Informatics, Medical Sciences, Natural Language Processing, Pathology, Pathology Informatics, Pathology Reports, Pediatric Appendicitis, Pediatric Appendicitis Score, Pediatric Medicine, Quality, Statistical Natural Language Processing, Text Classification, Text Mining, Unstructured Data

COMPUTATIONAL METHODS FOR CATEGORIZING UNSTRUCTURED DATA RELATED TO PEDIATRIC APPENDICITIS WITHIN ELECTRONIC MEDICAL RECORDS

by

BRITTANY NORMAN

B.A., The University of Georgia, 2011

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2017

COMPUTATIONAL METHODS FOR CATEGORIZING UNSTRUCTURED DATA RELATED TO PEDIATRIC APPENDICITIS WITHIN ELECTRONIC MEDICAL RECORDS

by

BRITTANY NORMAN

Major Professor:	Khaled Rasheed
Committee:	Shannon Quinn Walter Potter

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2017

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

The use of electronic medical records (EMR) is becoming the norm in the healthcare industry. In 2006, for example, only 46% of emergency departments in the United States used electronic records; by 2011 that number was 84% [1]. Now that so many medical data are available in electronic form, healthcare organizations are leveraging computational methods to analyze and extract insights from these data in order to improve patient care.

The healthcare data come in both structured (e.g. numerical) and unstructured forms (e.g., images, text). Unstructured medical data in the form of free text, such as physicians' notes/comments or pathology reports, provide additional challenges to those who want to collect and analyze relevant data using computational methods. Unlike numerical data, check-boxes, drop-down menus, or standardized data fields such as ICD-10 codes, when healthcare practitioners create natural language data they can describe a single concept in countless different ways. Furthermore, the large number of electronic medical records and the rarity of relevant records can make manual chart review infeasible.

To handle this kind of dataset, techniques from statistical natural language processing (NLP) can be applied to convert the data from an unstructured to a structured form (e.g., to numerical or categorical data).

Statistical NLP [13] is a proper subset of NLP which draws heavily from quantitative fields such as statistics, machine learning (ML), information theory, probability theory, and linear algebra.[1]

A wide variety of possible methods exist within the field of NLP, as well as within statistical NLP. Experimentation is required to determine which methods are best for the given domain (e.g. medical vs. retail vs. finance, etc.) and the given application case (e.g. appendicitis vs. influenza, etc.). Thus far, medical informaticists have successfully applied NLP to diverse projects such as extracting smoking status for asthma research [2], discovering adverse drug events [3], and identifying postoperative complications [4], among others.

In this thesis, automated processes for classifying unstructured health data related to pediatric appendicitis are explored. The project in Chapter 2 explores approaches to identifying electronic emergency department (ED) notes which contain features of Pediatric Appendicitis Score (PAS), such as right lower quadrant tenderness. The project in Chapter 3 compares procedures for detecting pathology reports which are positive for appendicitis.

The motivation behind the project in Chapter 2 was a quality analysis of physicians' use of PAS during their decision-making process. When PAS is less than or equal to 4, risk for appendicitis is low, and diagnostic imaging is not required. When PAS is greater than or equal to 8, risk for appendicitis is high, and consultation with surgery should begin instead of diagnostic imaging. Imaging should only be conducted when

[1] Symbolic NLP, on the other hand, draws heavily from the disciplines of linguistics and logic. Symbolic NLP also brings structure to the unstructured data for the purposes of analysis, but in different ways. One way is to use what is known about the sciences of human language and the human mind to detect structure; another way is to convert the complex human languages (i.e. natural languages) into simpler formal (i.e. invented) languages.

PAS is in the range from 5 to 7. Outside of this range, minimizing the amount of imaging is desired, both for efficiency purposes as well as for the purpose of minimizing the children's exposure to unnecessary radiation. The first step in this quality assessment was to determine which electronic medical records contained features of PAS. It is this first step which defines the scope of the project in Chapter 2.

The project in Chapter 3 was motivated by the quality department's interest in calculating the rate of negative appendicitis within the set of available appendectomy pathology reports. A low negative appendicitis rate is desirable, since this would indicate that very few appendectomies were performed when appendicitis was not actually present. The rate is calculated by taking the number of reports negative for appendicitis, then dividing by the number of appendectomy pathology reports minus the number of reports equivocal for appendicitis. It was infeasible to use human labor to calculate this rate, due to the large number of reports that would have to be read and classified by pathologists. Thus, the goal of the project in Chapter 3 was to develop an automated process to identify which pathology reports indicated negative appendicitis versus those which indicated positive.

CHAPTER 2

AUTOMATED IDENTIFICATION OF PEDIATRIC APPENDICITIS SCORE IN EMERGENCY DEPARTMENT NOTES USING NATURAL LANGUAGE PROCESSING[2]

[2] Norman, B., Davis, T., Quinn, S., Massey, R., and Hirsh, D. 2017. *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. 481-484. Reprinted here with permission of the publisher.

Abstract

Objective: The goal of this project was development of a software tool to detect documentation of Pediatric Appendicitis Score (PAS) within electronic emergency department (ED) notes. The overarching purpose was assessment of diagnostic imaging practices when PAS falls outside of a certain range, since minimizing patients' radiation exposure is desired.

Methods: 15074 ED notes were collected from visits between July 2011 – Aug. 2016. Notes were labeled as having PAS documented (PAS+) or not (PAS-). 12562 semistructured notes were split into 60% training, 20% validation, and 20% testing. An automated procedure was developed to label data, preprocess notes, extract features, construct three classification models, and compare the models. The selected model was also evaluated on a second testing set of 2512 hand-labeled (BN) unstructured notes using F1-score.

Results: The logistic regression (LR) model was selected for best F1-score on the validation set (0.9874). This model's F1-score on the human-labeled testing set of unstructured data (0.8391) outperformed the previous method (0.3435).

Discussion: The selected LR model demonstrated an improvement upon the previous method when evaluated on manually labeled unstructured data (no overlap in 95% CI).

Conclusion: While the LR classifier was trained and selected in an automated way, it still performed well compared to human performance. This tool can be used to expedite manual chart review for identification of PAS within ED notes.

Introduction

Background

As the use of electronic health records (EHR) becomes more prevalent—up from 46% of United States emergency departments in 2006 to 84% in 2011 [1]—healthcare organizations are leveraging computational methods to extract knowledge from these data. This wealth of electronic data provides numerous opportunities to improve patient care. The data come in both structured (e.g. numerical) and unstructured forms (e.g. text, images). Automated processing of unstructured textual data is a challenging task which requires techniques from natural language processing (NLP). Medical informaticists have successfully used NLP for applications such as extracting smoking status for asthma research [2], discovering adverse drug events [3], and identifying postoperative complications [4].

Objective

The goal of this project was the development of a software tool for detecting the documentation of Pediatric Appendicitis Score (PAS) [5] within emergency department (ED) notes (Fig. 2.1). If a physician performed PAS and documented this within an ED note, then the software should return that ED note to the end user. Note that the objective was not to develop a diagnosis system or a scoring system, but rather an information retrieval system.

The overarching goal was to improve quality of care by assessing the amount of diagnostic imaging being conducted when the PAS falls outside of a certain range. Due to the harmful effects of excessive exposure to radiation, imaging should be minimized. Imaging is not required when the PAS is too low (≤ 4) due to low suspicion for appendicitis, nor is it required when the PAS is too high (≥ 8) since high scores should lead to surgery consultation.

The previous approach to PAS detection used a regular expression to search for “Smart Phrases” which had been inserted into the ED notes by Epic Systems Corp. software users. These Smart Phrases are referred to as “semi-structured” text data, due to their predefined format (Fig. 2.2). While the regular expression (Fig. 2.3) does an excellent job of detecting these PAS Smart Phrases in the notes, it cannot detect PAS documentation that is entered in a free-form manner (Fig. 2.4). Thus, the objective was to develop an NLP system capable of detecting PAS documentation in completely unstructured text.

Methods

Data

This retrospective study was conducted using a collection of 15,074 electronic ED notes from two Children’s Healthcare of Atlanta locations, Scottish Rite and Egleston, with dates from July 12th 2011 to August 15th 2016. The vast majority of the visits included in the dataset were by children ranging from the ages of birth-18, however approximately 0.5% of visits were by patients over 18.

The datasets were prepared using a combination of computational and manual methods. The training and validation sets were collected and labeled in an automated manner. Two testing sets were collected: one semi-structured dataset which was labeled automatically, and one unstructured dataset which was hand-labeled (BN). The purpose was to train/validate the model without human intervention, then to test its performance against a human-labeled standard.

For the collection of semi-structured data (12,562 notes), the regular expression (RE) from Fig. 2.3 was used to automate the labeling process. The positive class (PAS+) was collected by selecting ED notes which had a PAS from 0-10 according to the RE pattern. The negative class (PAS-) was collected by taking a random sample of ED notes that did not match the RE pattern (i.e., PAS was null). Using stratified Bernoulli sampling, 7538 notes were designated for training (60%), 2512 notes were for validation (20%), and 2512 for testing (20%).

The second testing set of 2512 unstructured notes was collected by applying the model to unseen ED notes, then collecting two random samples, with 1256 of each classification type. These notes were then manually labeled (BN) as PAS- or PAS+ for a total of 1509 negative notes and 1003 positive notes.

The IRBs at both institutions were consulted. Both IRBs determined that the activity was an internal quality project and did not constitute research per the Federal human subject protection regulations. Thus, IRB review and approval were not required. Furthermore, the dataset was only accessible to Children's Healthcare of Atlanta employees with the relevant permissions.

Procedure

The software application was developed using a combination of NLP and machine learning (ML) methods (i.e. statistical NLP). The first phase consisted of standard NLP preprocessing steps, and the second phase used the resulting tokens as features to implement and train a classifier. Below is an overview of the procedure.

1) Preprocessing and Feature Extraction

    a) Cleaning

    b) Lowercasing

    c) Stop-word Removal

    d) Tokenizing

    e) TF-IDF

2) Model Construction

    a) Training

        i) Naïve Bayes

        ii) Support Vector Machine

        iii) Logistic Regression

    b) Validation

        i) F1-score to Compare 2.a.i-iii

    c) Testing

        i) F1-score of Best Model

Preprocessing and Feature Extraction

First, notes were cleaned by removing extraneous computer-generated text that was inserted upstream. This text included strings of repeated asterisks like "*****", HTML tags like "<BR>" and other auto-generated text such as "SMARTLIST_ METADATA_ BEGIN 7000009…". All words were lowercased so that features such as "Right" and "right" would be considered the same vocabulary feature. These preprocessed notes were tokenized into groups of 1-3 words: unigrams, bigrams, and trigrams. Also, the following 25 common stop words were removed from unigrams: *a, an, and, are, as, at, be, by, for,* *from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with* [6]. Lastly, the TF-IDF values (term frequency – inverse document frequency) were calculated for the remaining tokens, and the tokenized notes were converted into sparse TF-IDF vectors with 1,048,576 features.

Model Construction

Once notes are converted into TF-IDF vectors, they can be used as input into a wide variety of machine learning classifiers. The following kinds of supervised binary classifiers were implemented: Naïve Bayes, support vector machine, and logistic regression. The Naïve Bayes (NB) classifier used a smoothing parameter of 1.0. The linear support vector machine (SVM) used stochastic gradient descent (SGD) for training with a step size of 1.0 and L2 regularization with a parameter of 0.01. Training for the SVM ceases after either 100 iterations or convergence to 0.001. The logistic regression (LR) classifier was trained using the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) with 10 corrections used in the LBFGS update and L2 regularization with a parameter of 0.01. Training for the LR model stops after convergence to 0.0001 or 100 iterations.

All three classifiers were trained on the training set, then compared on the validation set using F1-score. The best-performing model on the validation set was selected and then evaluated on the two testing sets.

Results

Fig. 2.5 shows the performance of the three different classifiers on the validation set. (Scores are also shown for two other datasets which will be discussed shortly.) The LR classifier was chosen for highest performance (0.9874 F1-score) on this validation set. Recall that the validation set is computer-labeled, so this selection process is automated.

Once selected, the LR model was evaluated on the two testing sets (semistructured and unstructured) and compared with the RE approach (Fig. 2.5). Recall that the test sets are hand-labeled, so these scores are measured in comparison to human performance. Although the RE received the highest score (0.9980 F1) on semi-structured notes, it obtained the lowest score on unstructured notes (0.3435 F1). The LR model achieved the highest performance on unstructured notes (0.8391 F1). To see how these F-scores decompose into precision and recall, see Table 2.1.

Table 2.2 lists fifteen of the top features used by the LR model to identify positive cases of PAS documentation. These features all have an odds ratio that the class is PAS- which is less than or equal to 0.00004 (thus a high odds ratio that the class is PAS+). Revisit Fig. 2.2 to compare these features with an example of PAS documentation.

Discussion

While the LR classifier was trained and selected in a fully automated way on computer-labeled data, it was still able to perform well when compared to human performance using the test sets of manually-labeled examples. Furthermore, even though the training data were semi-structured due to the presence of Smart Phrases, the classifier was still able to perform well on the unstructured testing set data.

Most of the top features shown in Table 2.2 can be found in the example of PAS documentation from Fig. 2.2, thus a qualitative evaluation of these features is favorable. The features “abdominal” and “appetite” are interesting because even though they do not appear in the Smart Phrase from Fig. 2.2, the LR classifier still learned they were relevant to the concept of PAS. This is because these features appeared frequently enough in the PAS+ notes from the training set even though the features were located elsewhere in the note outside the bounds of the Smart Phrase.

Since the RE is better at detecting documentation of PAS within semi-structured Smart Phrases, and the NLP application is better at detecting PAS documentation within unstructured text, a hybrid approach could be used that combines the two methods to return relevant electronic medical records (EMR) to the end user.

Conclusion

A software tool was developed for the detection of PAS documentation within unstructured text from ED notes. The classification model (logistic regression) was trained and selected in a fully automated manner, including the data labeling process. The selected model performed well on a hand-labeled test set (0.8391 F1). This tool can be used to expedite manual chart review for the identification of PAS within electronic ED notes, and it demonstrates an improvement upon the existing computational method on unstructured data.

Future work could potentially include using similar methods to develop a tool for adult patients in addition to pediatric patients. The work could also be extended to include detection of other concepts/conditions in electronic physicians' notes.

Acknowledgements

We would like to thank the following people. From Children's Healthcare of Atlanta we thank: Vinay Boddula, Laura Gillard, Shervin Kharazmi, Lindsay Malenke, Myra Rolfes, and Mike Thompson. From the University of Georgia we thank: Qiang Hao, Bill

Hollingsworth, Fred Maier, Don Potter, Khaled Rasheed, Dylan Sims, and Ava Wright.

We also thank Jessica Sims.

References

[1] Jamoom, E., & Hing E. (2015). *Progress with electronic health record adoption among emergency and outpatient departments: United States, 2006–2011* (NCHS Data Brief No. 187). Hyattsville, MD: National Center for Health Statistics.

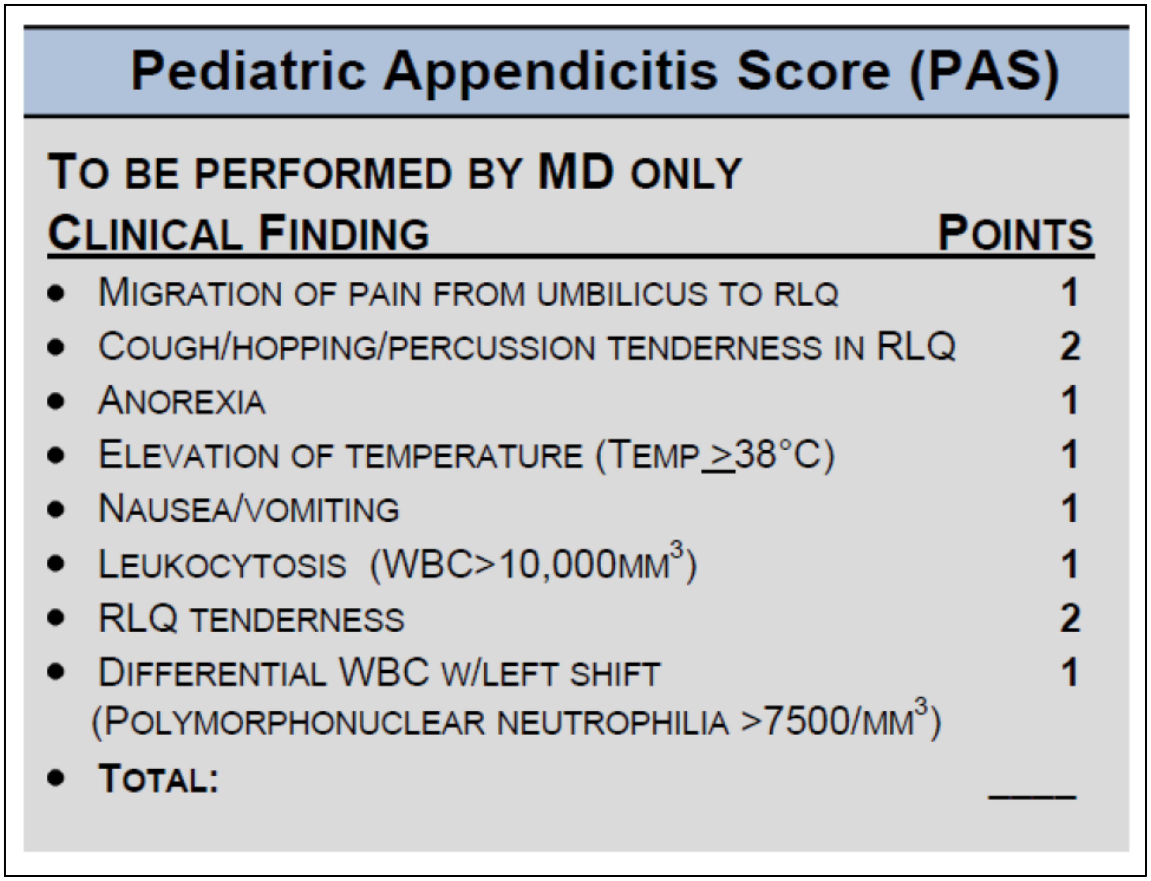
[2] Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N., & Lazarus, R. (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6(1), 30-39.

[3] Friedman, C. (2009). Discovering novel adverse drug events using natural language processing and mining of the electronic health record. In C. Combi, Y. Shahar, & A. Abu-Hanna (Eds.), *Lecture Notes in Computer Science Vol 5651* (pp. 1-5). Berlin: Springer-Verlag Berlin Heidelberg.

[4] Murff, H. J., FitzHenry, F., Matheny, M. E., Gentry, N., Kotter, K. L., Crimin, K., Dittus, R.S., Rosen, A.K., Elkin, P.L., Brown, S.H., & Speroff, T. (2011). Automated identification of postoperative complications within an electronic medical record using natural language processing. *Journal of the American Medical Association*, 306(8), 848-855.

[5] Samuel, M. (2002). Pediatric appendicitis score. *Journal of Pediatric Surgery*, 37(6), 877-881.

[6] Manning, C.D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. New York, NY: Cambridge University Press.

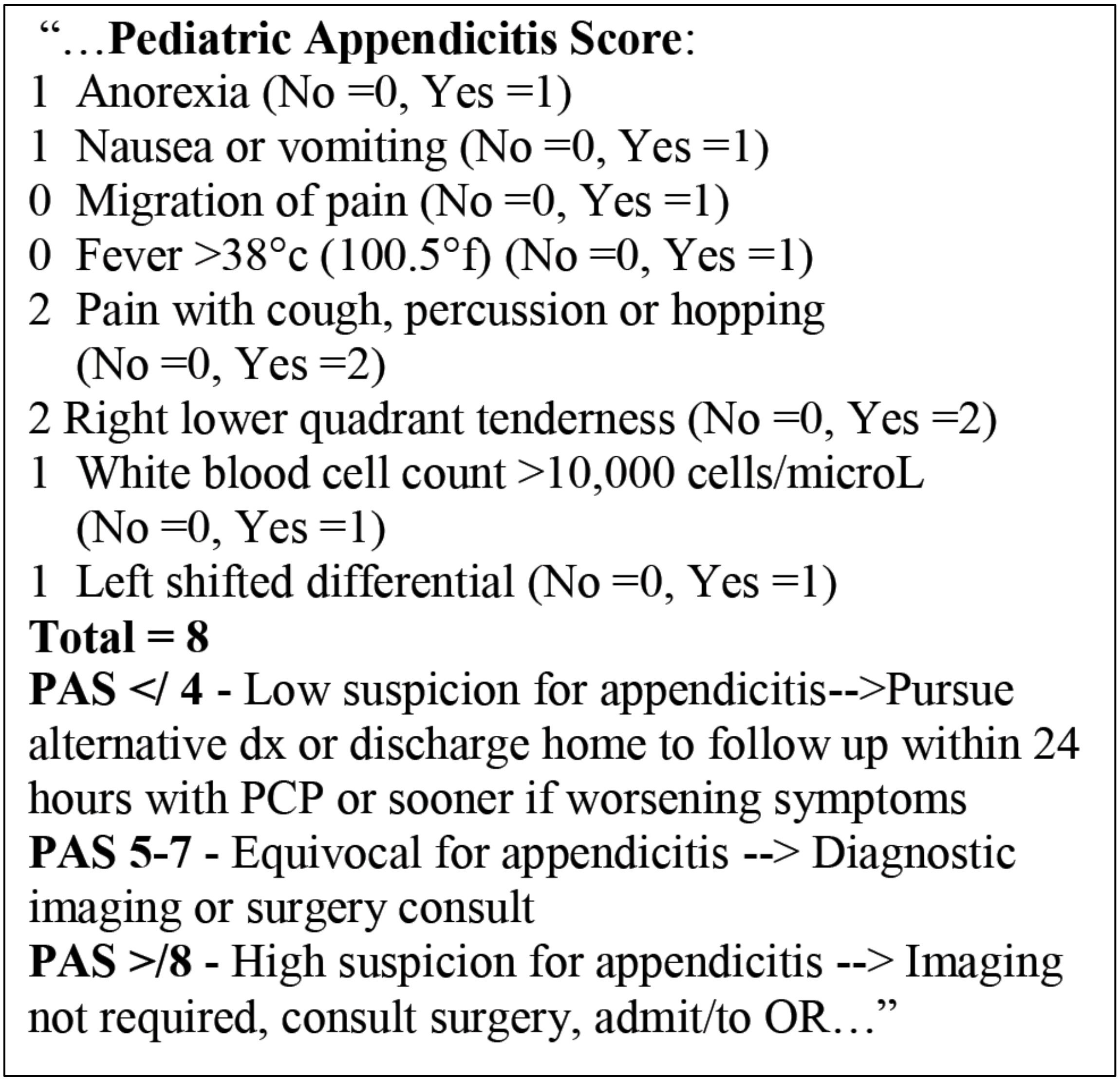


Pediatric Appendicitis Score (PAS)

To be performed by MD only

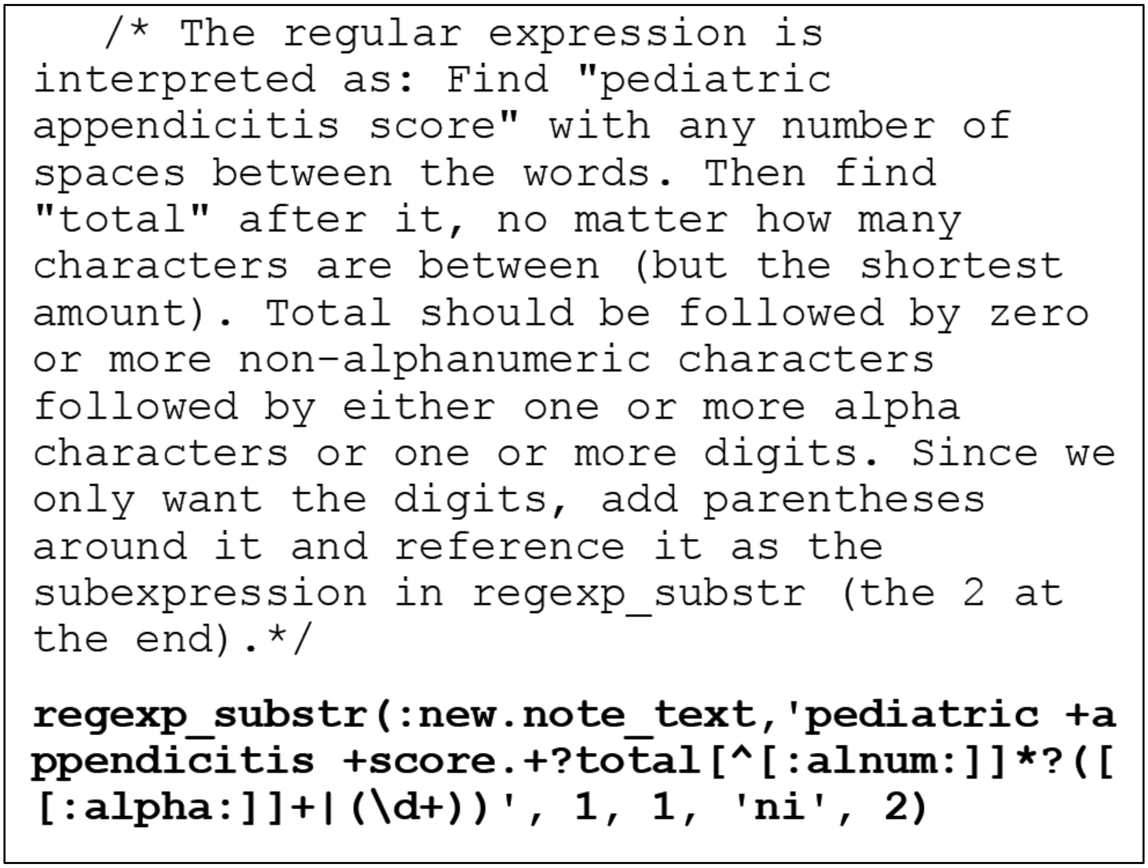
Clinical Finding	Points
• Migration of pain from umbilicus to RLQ	1
• Cough/hopping/percussion tenderness in RLQ	2
• Anorexia	1
• Elevation of temperature (Temp ≥38°C)	1
• Nausea/vomiting	1
• Leukocytosis (WBC>10,000mm^3)	1
• RLQ tenderness	2
• Differential WBC w/left shift (Polymorphonuclear neutrophilia >7500/mm^3)	1
• **Total:**	____

Figure 2.1: Pediatric Appendicitis Score (PAS). Illustration of how the PAS is calculated. Total can range from 0 to 10 inclusive.



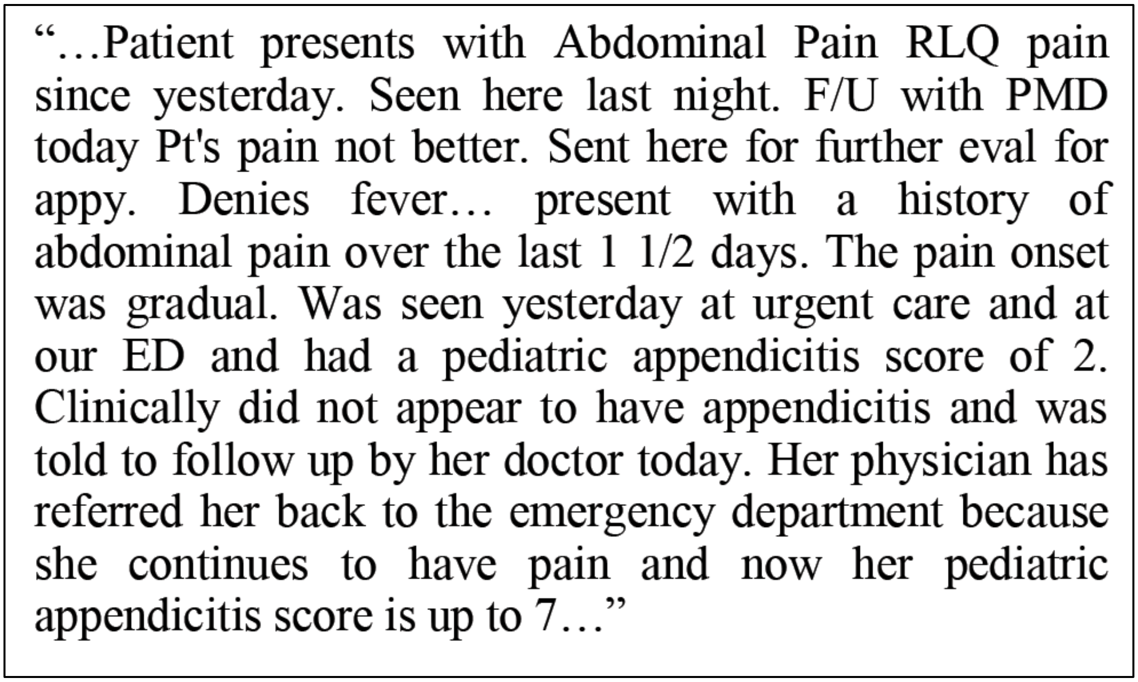
“…**Pediatric Appendicitis Score**:
1 Anorexia (No =0, Yes =1)
1 Nausea or vomiting (No =0, Yes =1)
0 Migration of pain (No =0, Yes =1)
0 Fever >38°c (100.5°f) (No =0, Yes =1)
2 Pain with cough, percussion or hopping (No =0, Yes =2)
2 Right lower quadrant tenderness (No =0, Yes =2)
1 White blood cell count >10,000 cells/microL (No =0, Yes =1)
1 Left shifted differential (No =0, Yes =1)
Total = 8
PAS </ 4 - Low suspicion for appendicitis**-->**Pursue alternative dx or discharge home to follow up within 24 hours with PCP or sooner if worsening symptoms
PAS 5-7 - Equivocal for appendicitis --> Diagnostic imaging or surgery consult
PAS >/8 - High suspicion for appendicitis **-->** Imaging not required, consult surgery, admit/to OR…”

Figure 2.2: PAS Smart Phrase. Example of text with a Smart Phrase. The MD only fills in the numbers for points and total.



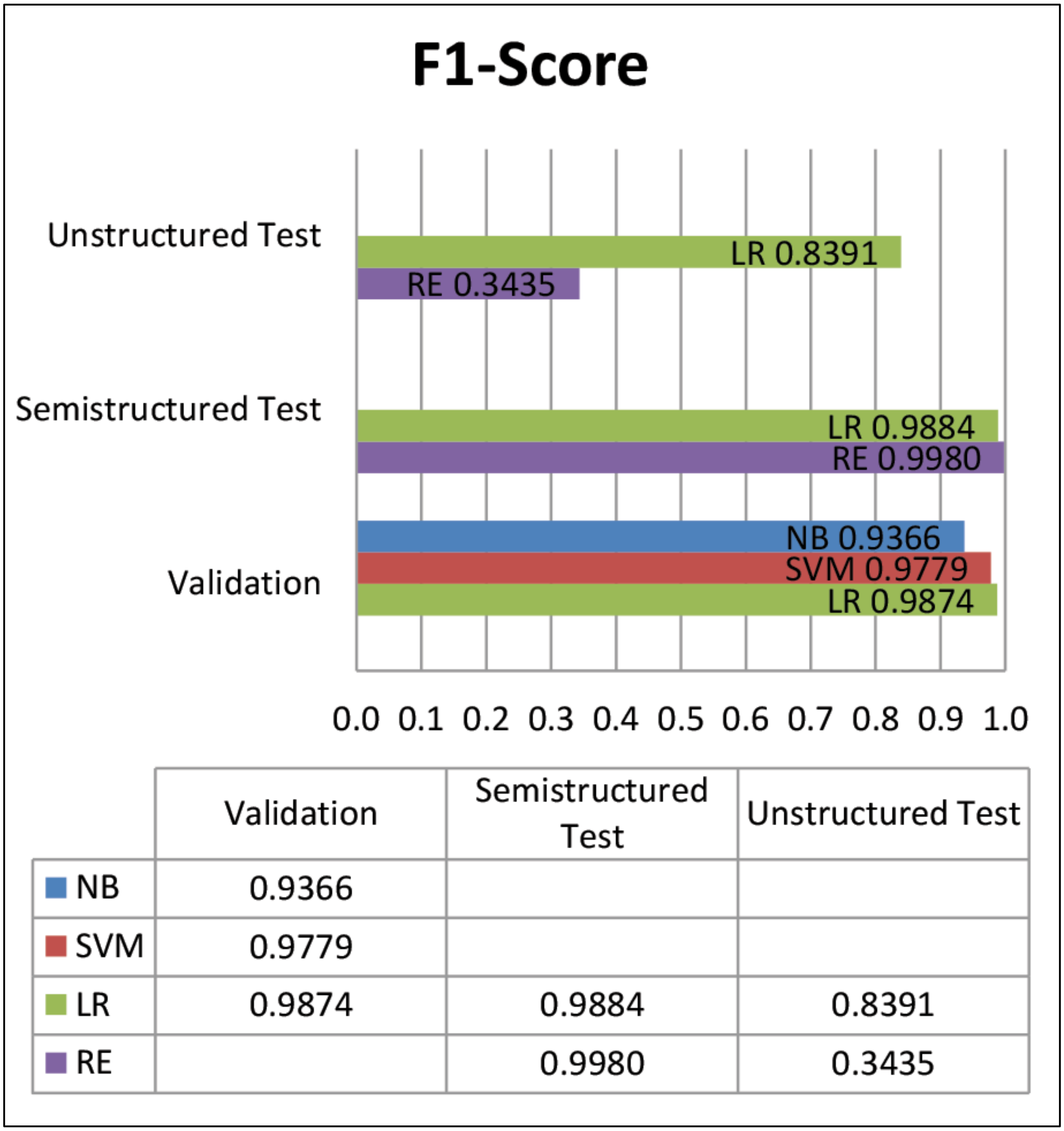
```
   /* The regular expression is
interpreted as: Find "pediatric
appendicitis score" with any number of
spaces between the words. Then find
"total" after it, no matter how many
characters are between (but the shortest
amount). Total should be followed by zero
or more non-alphanumeric characters
followed by either one or more alpha
characters or one or more digits. Since we
only want the digits, add parentheses
around it and reference it as the
subexpression in regexp_substr (the 2 at
the end).*/

regexp_substr(:new.note_text,'pediatric +a
ppendicitis +score.+?total[^[:alnum:]]*?([
[:alpha:]]+|(\d+))', 1, 1, 'ni', 2)
```

Figure 2.3: Regular Expression (RE). This RE extracts the PAS which were documented by MD inserting Smart Phrase. RE written in Oracle. (Note: Uses Posix character classes.)



“…Patient presents with Abdominal Pain RLQ pain since yesterday. Seen here last night. F/U with PMD today Pt's pain not better. Sent here for further eval for appy. Denies fever… present with a history of abdominal pain over the last 1 1/2 days. The pain onset was gradual. Was seen yesterday at urgent care and at our ED and had a pediatric appendicitis score of 2. Clinically did not appear to have appendicitis and was told to follow up by her doctor today. Her physician has referred her back to the emergency department because she continues to have pain and now her pediatric appendicitis score is up to 7…”

Figure 2.4: Free-Form PAS Documentation. Here is one example of what documentation of PAS could look like in unstructured text. The possibilities are innumerable.



	Validation	Semistructured Test	Unstructured Test
🟦 NB	0.9366		
🟥 SVM	0.9779		
🟩 LR	0.9874	0.9884	0.8391
🟪 RE		0.9980	0.3435

Figure 2.5: F1-Score Performance. All numerical values represent F1-scores. NB=Naïve Bayes. SVM=Support Vector Machine. LR=Logistic Regression. RE=Regular Expression.

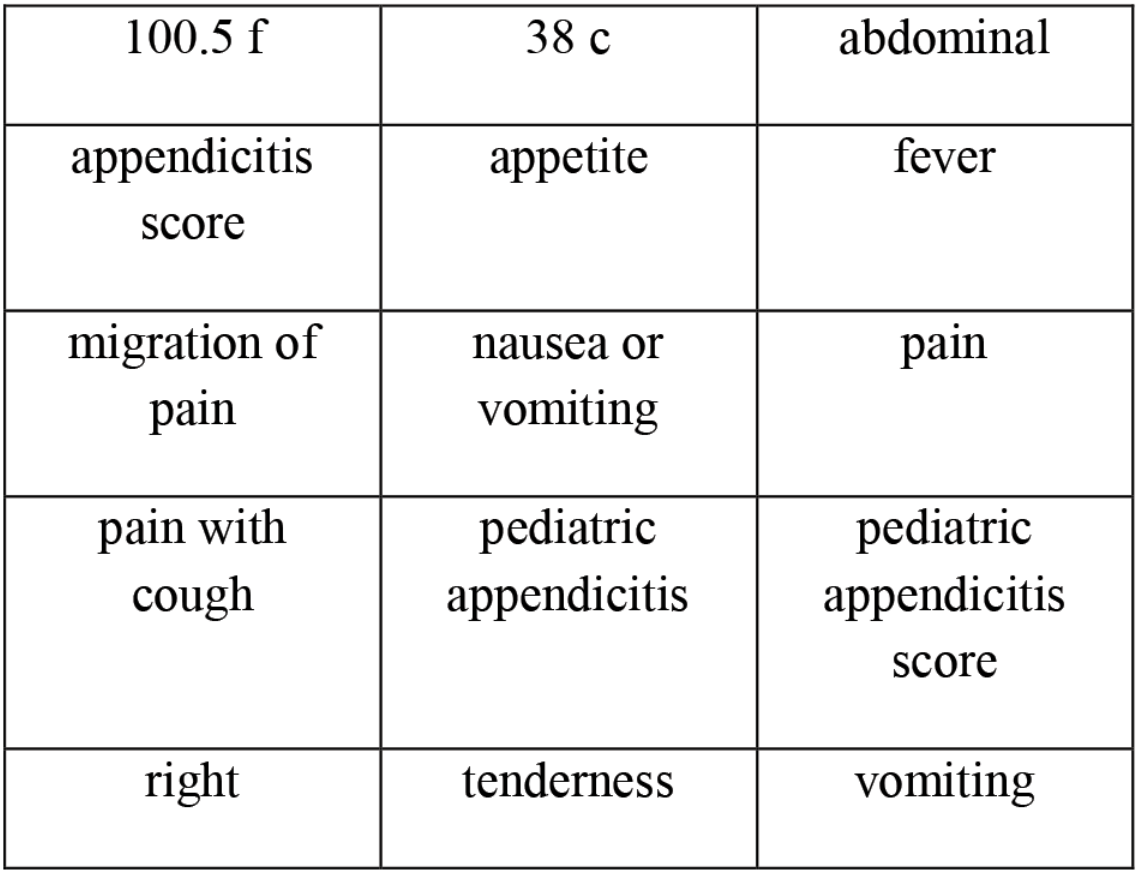
TABLE 2.1: PRECISION AND RECALL

Validation Set			
	NB	SVM	**LR**
F1-Score	0.9366	0.9779	**0.9874**
Precision	0.9119	**0.9878**	0.9767
Recall	0.9626	0.9682	**0.9984**

Semi-Structured Test Set		
	RE	LR
F1-Score	**0.9980**	0.9884
Precision	**0.9976**	0.9801
Recall	**0.9984**	0.9969

Unstructured Test Set		
	RE	**LR**
F1-Score	0.3435	**0.8391**
Precision	**0.9999**	0.7488
Recall	0.2074	**0.9541**

TABLE 2.2: FIFTEEN TOP FEATURES



100.5 f	38 c	abdominal
appendicitis score	appetite	fever
migration of pain	nausea or vomiting	pain
pain with cough	pediatric appendicitis	pediatric appendicitis score
right	tenderness	vomiting

CHAPTER 3

AUTOMATING THE DETECTION OF NEGATIVE APPENDICITIS IN PATHOLOGY REPORTS USING NATURAL LANGUAGE PROCESSING[3]

[3] Norman, B., Davis, T., Gillard, L., Massey, R., and Hirsh, D. To be submitted to 2018 *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI).*

Abstract

The motivation behind this project was to calculate “negative appendicitis rate” with respect to appendectomy pathology reports, for quality assessment purposes at two hospitals. This project’s task was the development of a computer software tool to discriminate between appendectomy pathology reports negative for appendicitis vs. positive.

The dataset included 8595 pathology reports (8019 positive, 576 negative). Reports were collected from two children’s hospitals between 06/19/2014 and 11/15/2016 for patients at least 5 years old at the start of the visit. A statistical natural language processing (NLP) approach was used to build the classification model. The pathology reports were converted from unstructured text into machine-readable data using cleaning, preprocessing, and feature extraction. Numerical features were extracted using TF-IDF (term frequency – inverse document frequency) for a total of 1,048,576 possible features. Several classification models were compared: logistic regression (LR), Naïve Bayes (NB), support vector machine (SVM), decision tree (DT), random forest (RF), and gradient-boosted tree (GBT). Some model candidates used a smaller number of features (1000). All models were built with the training data (5157, 60%) and compared with the validation data (1719, 20%) using the F-Score metric. The final selected model was evaluated with the testing data (1719, 20%).

The logistic regression model with 1000 features was selected due to highest performance on the validation set. This model received a 0.9960 F-Score when evaluated on the testing data. With 95% confidence, the expected F-Score of the chosen model should range between 0.9930 and 0.9990.

Introduction

Goals

The quality department was interested in knowing the “negative appendicitis rate” with respect to the appendectomy pathology reports. A low rate is desired, as this indicates that only a few appendectomies are being performed when appendicitis is not actually present. This rate is calculated by taking the number of reports which are negative for appendicitis and dividing by the number of visits with an appendectomy where an appendix pathology report is present (minus the number of reports which are equivocal for appendicitis). Due to the large number of reports, it was not feasible to have a human calculate this rate. Thus, the goal was to develop automated methods for the calculation by developing a computer software tool.

The challenge in automating this process was due to the fact that the pathology reports consisted of unstructured textual data, and pathologists were using many different ways to describe positive appendicitis. For example, a pathologist could use any of the phrases in Fig 3.1 to indicate the positive case. Similarly, a variety of phrases (See Fig 3.2) could be used to indicate negative appendicitis.

The previous method for automatically calculating the negative appendicitis rate involved hand-crafting a set of rules, guided by a pathologist expert, and implementing these rules using regular expressions to search for particular substrings. The time-consuming and labor-intensive nature of this method led the department to seek an alternative approach. Thus, the purpose of this project was to use natural language processing to develop an alternative method for automation, then to compare it with the previous method.

Background

Categorizing pathology reports using automated methods can be a challenging task. This is due to the unstructured nature of the reports and the variety of ways that a pathologist can describe the same concept. In a study of breast pathology reports, for instance, pathologists used 124 different ways of saying the concept invasive ductal carcinoma, 95 ways of saying invasive lobular carcinoma, and over 4000 ways of saying that invasive ductal carcinoma was not present [7]. When there is potential for such a wide variation in terminology, it can be difficult for pathologists to bring to mind all of the possibilities in order to compile an exhaustive list to be used for the automated process.

An alternative to compiling an exhaustive list of phrases is to use techniques from statistical natural language processing, which draws from the disciplines of natural language processing (NLP), statistics, and machine learning (ML). First, NLP can be used to convert the unstructured data into a machine-readable form. Then, ML or statistical methods can be used to learn which phrases are relevant to the concept at hand and to build a model that can classify new pathology reports accordingly. These techniques have been successfully used to categorize various kinds of pathology reports, such as breast reports [7], colonoscopy reports [8] and those for liver cancer [9], prostate cancer [10], and other kinds of cancer [11].

Methods

Data

The dataset consisted of 8595 appendectomy pathology reports, 8019 of which were labeled positive for appendicitis and 576 of which were labeled negative. The reports were collected from two Children's Healthcare of Atlanta hospitals (Egleston and

Scottish Rite) between 06/19/2014 and 11/15/2016 for patients who were at least 5 years old at the start of the appendectomy visit. These reports were labeled using a hand-written (LG) series of rules, implemented using regular expressions. The rules proceed as follows:

1. First remove the following types of reports:
    a. Remove reports that do not contain substring “append” (case-insensitive).
    b. Remove reports with multiple specimens.
    c. Remove reports with incidental appendectomies or where appendix is not identified.
2. Separate remaining reports into those that do or do not mention appendicitis.
3. For the reports that mention appendicitis:
    a. Classify as negative if phrase is immediately preceded by a negating statement (Fig. 3.3).
    b. Otherwise, classify as positive.
4. For the reports that do not mention appendicitis:
    a. Classify as positive if report includes a phrase indicating positive appendicitis (Fig. 3.1).
    b. Otherwise, classify reports as follows:
        i. Classify as negative if report includes a relevant negating statement (Fig. 3.4).
        ii. Otherwise, classify as equivocal/unknown

The IRBs were consulted at both institutions involved in the project. Both IRBs decided that the project did not constitute research per the Federal human subject protection regulations and that the activity was an internal quality project. As a result, review and approval by the IRB were not required. In addition, only Children's Healthcare of Atlanta employees with the relevant permissions were given access to the dataset.

Procedure

Overview

The input data were used to learn a computational model that can classify new unseen pathology reports as negative or positive for appendicitis. Several kinds of models were considered and compared before selecting the final model; these include logistic regression (LR), Naïve Bayes (NB), support vector machine (SVM), decision tree (DT), random forest (RF), and gradient-boosted tree (GBT).

Before building and comparing these six kinds of machine-learning models, the pathology reports were converted from unstructured, free-form text into a more structured format: a matrix of data with each row representing an individual report, each column representing different features (i.e. words or groups of words) found in the reports, and the last column representing the type of report (positive or negative). Before converting the dataset into a matrix, the reports were cleaned and preprocessed. TF-IDF values (term frequency – inverse document frequency) were then calculated and used to populate the cells of the matrix.

Cleaning, Preprocessing, and Feature Extraction

The pathology reports were cleaned of extraneous text that had been introduced somewhere upstream. Each report is tokenized into groups of one, two, and three words. Stop-words were removed (*a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its,* *of, on, that, the, to, was, were, will, with* [12]) from the list of unigrams due to their lack of value as features. All text was lowercased so that the strings “appendicitis”, “Appendicitis”, and “APPENDICITIS”, for example, were considered the same feature. Finally, TF-IDF values were calculated. There were a total of 1,048,576 unique features, and some of the models used all of these features while others used a smaller number (1000).

Building and Comparing the Models

The LR model, NB model, and SVM model were built using the full feature set. The DT, RF, and GBT models were built using a smaller number of features (1000) because training decision-tree-based models on the full set was inefficient. These 1000 features were created using the feature mixing method [13]. The dataset was split into stratified sets of 60% training (5157), 20% validation (1719), and 20% testing (1719). All models were compared on the validation set, and the best model was chosen for final evaluation on the testing set.

The following hyperparameters were used for the models. The NB models used a smoothing parameter of 1.0 to avoid the loss of information that could occur if probabilities were instead being multiplied by zero. The LR models used the L-BFGS algorithm (limited-memory Broyden Fletcher Goldfarb Shanno) with 10 corrections used in the L-BFGS update, and stopped training after reaching a convergence tolerance of 1×10^{-4} or upon reaching 100 iterations. L2-regularization was used with a regularizer parameter of 0.01. The linear SVM models used the SGD algorithm (stochastic gradient descent) for training with a step size of 1.0. Training ceased after reaching the convergence tolerance of 0.001 or after completing 100 iterations. L2-regularization was used with a regularizer parameter of 0.01.

The following hyperparameters were used for the decision trees and ensembles of trees. For the DT model, entropy was used to measure impurity and compare features. The minimum number of instances required at the child nodes to create a parent split was set to 1, and splits were only allowed if there was an information gain of 0.0 or greater. The number of bins used for finding splits at each node was 32. Lastly, the maximum allowable depth for the tree was 5 levels. For the RF model, 3 trees were used in building the ensemble, and all parameters for each of the 3 trees were the same as those in the DT model. For the GBT model, 3 trees were used. Log loss was used for minimization during the gradient boosting with 100 iterations and a learning rate of 0.1. For each of the 3 trees in the GBT model, the same parameters were used as for the DT model.

Results

All nine models were compared on the validation set using F-Score. These included the LR, NB, and SVM models using the full feature set (1,048,576), and the LR, NB, SVM, DT, RF, and GBT models using the smaller feature set (1000). The results of this comparison can be seen in Fig. 3.5 and Table 3.1. The logistic regression model with 1000 features and the random forest model with 1000 features tied for the highest score on the validation set. Of these two models, the logistic regression model was chosen due to its greater simplicity. The chosen logistic regression model's performance on the final testing set was an F-Score of 0.9960 (Table 3.2).

Discussion

With 95% confidence, the expected F-Score of the chosen logistic regression model should range between 0.9930 and 0.9990, based on the model's performance on the testing set. Since the testing data were labeled in an automated fashion, one should keep in mind that these F-Scores are relative to a computer-labeled gold standard. In the future, this project will be extended by evaluating the same methodology on a hand-labeled test set. Since the new testing set will be manually labeled by a pathologist, it will thus provide a basis for comparison with human performance on the same task (discrimination between appendectomy pathology reports which are negative for appendicitis versus positive). This will provide a means to compare the existing computational method (hand-crafted rules implemented with regular expressions) to the new computational method (natural language processing) using human expertise as the gold standard.

Conclusion

A computational model was developed to automate the classification of appendectomy pathology reports into negative vs. positive for appendicitis using natural language processing techniques. The logistic regression model achieved an F-Score of 0.9960. After the model is evaluated on a pathologist-labeled testing set in the future, there is potential for the tool to be usable for expediting or replacing manual chart review if performance levels are promising. To extend this work, methods similar to those described in this paper could be attempted in the case of adult appendectomies or other kinds of pathology reports.

Acknowledgements

From Children's Healthcare of Atlanta, we would like to thank: Shervin Kharazmi, Lindsay Malenke, Myra Rolfes, and Mike Thompson. From the University of Georgia, we would like to thank: Vinay Boddula, Qiang Hao, Bill Hollingsworth, Fred Maier, Don Potter, Shannon Quinn, Khaled Rasheed, Dylan Sims, and Ava Wright. We would also like to thank Jessica Sims.

References

[7] Buckley, J. M., Coopey, S. B., Sharko, J., Polubriaginof, F., Drohan, B., Belli, A. K., Kim, E.M.H., Garber, J.E., Smith, B.L., Gadd, M.A., Specht, M. C., Roche, C.A., Gudewicz, T.M., Hughes, K.S. (2012). The feasibility of using natural language processing to extract clinical information from breast pathology reports. *Journal of pathology informatics*, 3(1), 23.

[8] Imler, T. D., Morea, J., Kahi, C., & Imperiale, T. F. (2013). Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clinical Gastroenterology and Hepatology*, 11(6), 689–694.

[9] Sada, Y., Hou, J., Richardson, P., El-Serag, H., & Davila, J. (2016). Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. *Medical care*, 54(2), e9-e14.

[10] Thomas, A. A., Zheng, C., Jung, H., Chang, A., Kim, B., Gelfond, J., Slezak, J., Porter, K., Jacobsen, S.J, & Chien, G. W. (2014). Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. *World journal of urology*, 32(1), 99-103.

[11] Nguyen, A., Moore, J., Zuccon, G., Lawley, M., & Colquist, S. (2012). Classification of pathology reports for cancer registry notifications. In *Health Informatics: Building a Healthcare Future Through Trusted Information: Selected Papers from the 20th Australian National Health Informatics Conference (HIC 2012)* (pp. 150-156). IOS Press.

[12] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* NewYork: Cambridge University Press.

[13] Ganchev, K., & Dredze, M. (2008). Small statistical models by random feature mixing. Proceedings from ACL '08: *Association for Computational Linguistics HLT (Human Language Technology) Workshop on Mobile Language Processing*, 19-20.

[14] Manning, C. D., & Schütze, H. (2003). *Foundations of statistical natural language processing*. MIT Press.

[15] Mitchell, T. M. (2013). *Machine learning*. McGraw Hill Education.

[16] Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Upper Saddle River, New Jersey: Prentice Hall.

[17] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

“FECALITH”
“MUCOSITIS”
“SEROSITIS”
“ENTEROBIUS VERMICULARIS”
“MUCOSAL ACUTE INFLAMMATION”
“ACUTE MUCOSAL INFLAMMATION”
“CHRONIC INFLAMMATION OF SEROSA”
“FOCAL ACUTE INFLAMMATION”
“SUPPURATIVE INFLAMMATION”
“FIBROUS OBLITERATION”
“SEROSAL ADHESIONS”
“VASCULAR CONGESTION”
“TUMOR”
“MELANOSIS”
“DISTAL LUMINAL OBLITERATION”
“MUCOSAL EROSION”
“CONGESTED SEROSAL”
“CHRONIC INFLAMMATORY”
“MURAL CONGESTION”
“ACUTE MUCOSAL CONGESTION”
“NEUTROPHILIC CRYPTITIS”
“MUCOSAL INFLAMMATION”
“FOCAL NEUTROPHILIC CRYPT ABSCESS”
“MURAL INFLAMMATORY CHANGES”
“MINIMAL MUCOSAL INFLAMMATORY CHANGES”

Figure 3.1: Phrases Indicating Positive for Appendicitis.

“NO EVIDENCE OF ACUTE APPENDICITIS”
“NO EVIDENCE OF APPENDICITIS”
“NO DIAGNOSTIC ABNORMALIT”
“NO DEFINITE EVIDENCE OF ACUTE APPENDICITIS”
“NO EVIDENCE OF TRANSMURAL INVOLVEMENT”
“NO SIGNIFICANT INFLAMMATION”
“NO DIAGNOSTIC ABNORMALIT”
“NO PATHOLOGIC ABNORMALIT”
“NO SIGNIFICANT ABNORMALIT”
“NO HISTOPATHOLOGIC ALTERATION”
“NO TRANSMURAL INFLAMMATION”
“NO EVIDENCE OF TRANSMURAL INFLAMMATION”
“NO SIGNIFICANT PATHOLOGIC ABNORMALIT”
“NO ACUTE INFLAMMATION”
“NO SIGNIFICANT ACUTE INFLAMMATION”
“NO INFLAMMATORY CHANGES IDENTIFIED”
“NO HISTOPATHOLOGIC ABNORMALIT”
“NO SIGNIFICANT INFLAMMATION”
“WITHOUT SIGNIFICANT PATHOLOGIC CHANGES”
“NO INFLAMMATORY CHANGES”
“NO SIGNIFICANT INFLAMMATION”
“NO HISTOPATHOLOGIC ABNORMALITY”
“NO SIGNIFICANT PATHOLOGIC CHANGE”
“WITHOUT SIGNIFICANT DIAGNOSTIC ABNORMALITY”

Figure 3.2: Phrases Indicating Negative for Appendicitis

"NO EVIDENCE OF ACUTE APPENDICITIS"

"NO EVIDENCE OF APPENDICITIS"

"NO DIAGNOSTIC ABNORMALIT"

"NO DEFINITE EVIDENCE OF ACUTE APPENDICITIS"

"NO EVIDENCE OF TRANSMURAL INVOLVEMENT"

"NO SIGNIFICANT INFLAMMATION"

Figure 3.3: Negative Label Phrases

"NO DIAGNOSTIC ABNORMALIT"

"NO PATHOLOGIC ABNORMALIT"

"NO SIGNIFICANT ABNORMALIT"

"NO HISTOPATHOLOGIC ALTERATION"

"NO TRANSMURAL INFLAMMATION"

"NO EVIDENCE OF TRANSMURAL INFLAMMATION"

"NO SIGNIFICANT PATHOLOGIC ABNORMALIT"

"NO ACUTE INFLAMMATION"

"NO SIGNIFICANT ACUTE INFLAMMATION"

"NO INFLAMMATORY CHANGES IDENTIFIED"

"NO HISTOPATHOLOGIC ABNORMALIT"

"NO SIGNIFICANT INFLAMMATION"

"WITHOUT SIGNIFICANT PATHOLOGIC CHANGES"

"NO INFLAMMATORY CHANGES"

"NO SIGNIFICANT INFLAMMATION"

"NO HISTOPATHOLOGIC ABNORMALITY"

"NO SIGNIFICANT PATHOLOGIC CHANGE"

"WITHOUT SIGNIFICANT DIAGNOSTIC ABNORMALITY"

Figure 3.4: Additional Negative Label Phrases

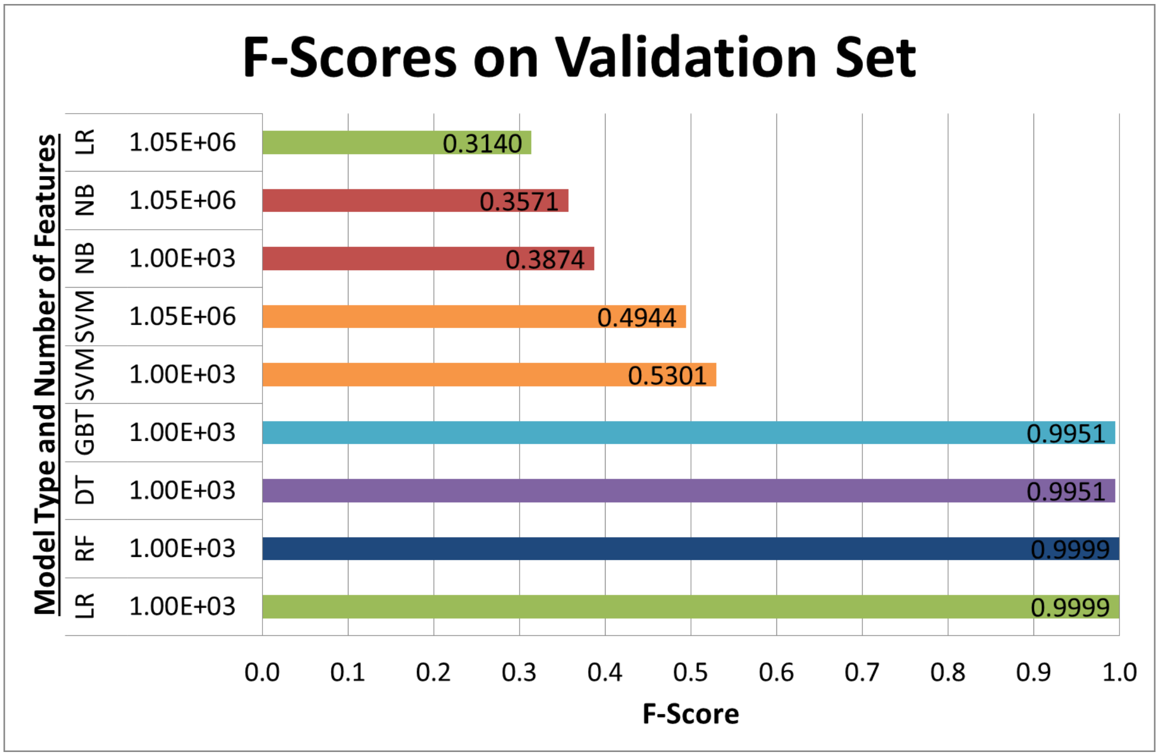


Figure 3.5: F-Scores on Validation Set. LR=Logistic Regression, NB=Naïve Bayes, SVM=Support Vector Machine, GBT= Gradient Boosted Trees, DT=Decision Tree, and RF=Random Forest.

TABLE 3.1: F-SCORES ON VALIDATION SET

Name	Number of Features	F-Score
LR	1.00E+03	0.9999
RF	1.00E+03	0.9999
DT	1.00E+03	0.9951
GBT	1.00E+03	0.9951
SVM	1.00E+03	0.5301
SVM	1.05E+06	0.4944
NB	1.00E+03	0.3874
NB	1.05E+06	0.3571
LR	1.05E+06	0.3140

TABLE 3.2: TESTING SET PERFORMANCE

Name	Logistic Regression
Number of Features	1.00E+03
Precision	0.9999
Recall	0.9920
F-Score	0.9960

CHAPTER 4

CONCLUSION

Computational methods from statistical natural language processing (NLP) were used to develop two software tools for categorizing unstructured data in electronic medical records. Both tools categorized data related to pediatric appendicitis; the first project was built to detect features of Pediatric Appendicitis Score (PAS) within emergency department (ED) notes; the second project was built to identify negative appendicitis within appendectomy pathology reports.

The final model chosen in Chapter 2 to detect PAS features in ED notes was a logistic regression model with 1,048,576 features. The model achieved an F-Score of 0.8391 when compared to human performance. With 95% confidence, expected performance of this model should range between 0.8247 to 0.8535. This model improves upon the previous computational method (0.3435 F-Score, 95% CI [0.3249,0.3621]). This new method could be used to expedite the chart review process by narrowing down the number of charts to be reviewed. In the future, performance of this model could be further improved by incorporating negation detection into the feature extraction process.

In Chapter 3, the chosen model to identify negative appendicitis in pathology reports was a logistic regression model with 1000 features which were created using the feature mixing technique. The F-Score for this model was 0.9960, with a 95% confidence interval between 0.9930 and 0.9990. Since the testing set was labeled automatically instead of manually, this performance metric is relative to a computer-labeled gold standard, not a human one. Thus, in the future, a manually reviewed dataset labeled by pathologists could provide a basis for comparison between the previous computational technique and the new NLP technique.

The methods described in this thesis have been applied to categorizing unstructured data related to pediatric appendicitis. In the future, this work could be extended by applying similar techniques to adult appendicitis, other types of pathology reports beyond appendectomy reports, or other types of unstructured data found within electronic medical records.

REFERENCES

[1] Jamoom, E., & Hing E. (2015). *Progress with electronic health record adoption among emergency and outpatient departments: United States, 2006–2011* (NCHS Data Brief No. 187). Hyattsville, MD: National Center for Health Statistics.

[2] Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N., & Lazarus, R. (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6(1), 30-39.

[3] Friedman, C. (2009). Discovering novel adverse drug events using natural language processing and mining of the electronic health record. In C. Combi, Y. Shahar, & A. Abu-Hanna (Eds.), *Lecture Notes in Computer Science Vol 5651* (pp. 1-5). Berlin: Springer-Verlag Berlin Heidelberg.

[4] Murff, H. J., FitzHenry, F., Matheny, M. E., Gentry, N., Kotter, K. L., Crimin, K., Dittus, R.S., Rosen, A.K., Elkin, P.L., Brown, S.H., & Speroff, T. (2011). Automated identification of postoperative complications within an electronic medical record using natural language processing. *Journal of the American Medical Association*, 306(8), 848-855.

[5] Samuel, M. (2002). Pediatric appendicitis score. *Journal of Pediatric Surgery*, 37(6), 877-881.

[6] Manning, C.D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. New York, NY: Cambridge University Press.

[7] Buckley, J. M., Coopey, S. B., Sharko, J., Polubriaginof, F., Drohan, B., Belli, A. K., Kim, E.M.H., Garber, J.E., Smith, B.L., Gadd, M.A., Specht, M. C., Roche, C.A., Gudewicz, T.M., Hughes, K.S. (2012). The feasibility of using natural language processing to extract clinical information from breast pathology reports. *Journal of pathology informatics*, 3(1), 23.

[8] Imler, T. D., Morea, J., Kahi, C., & Imperiale, T. F. (2013). Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clinical Gastroenterology and Hepatology*, 11(6), 689–694.

[9] Sada, Y., Hou, J., Richardson, P., El-Serag, H., & Davila, J. (2016). Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. *Medical care*, 54(2), e9-e14.

[10] Thomas, A. A., Zheng, C., Jung, H., Chang, A., Kim, B., Gelfond, J., Slezak, J., Porter, K., Jacobsen, S.J, & Chien, G. W. (2014). Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. *World journal of urology*, 32(1), 99-103.

[11] Nguyen, A., Moore, J., Zuccon, G., Lawley, M., & Colquist, S. (2012). Classification of pathology reports for cancer registry notifications. In *Health Informatics: Building a Healthcare Future Through Trusted Information: Selected Papers from the 20th Australian National Health Informatics Conference (HIC 2012)* (pp. 150-156). IOS Press.

[12] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* NewYork: Cambridge University Press.

[13] Ganchev, K., & Dredze, M. (2008). Small statistical models by random feature mixing. Proceedings from ACL '08: *Association for Computational Linguistics HLT (Human Language Technology) Workshop on Mobile Language Processing*, 19-20.

[14] Manning, C. D., & Schütze, H. (2003). *Foundations of statistical natural language processing*. MIT Press.

[15] Mitchell, T. M. (2013). *Machine learning*. McGraw Hill Education.

[16] Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Upper Saddle River, New Jersey: Prentice Hall.

[17] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques.* Morgan Kaufmann.