Automated Identification of Pediatric Appendicitis Score in Emergency Department Notes using Natural Language Processing*

Brittany Norman, Tod Davis, Shannon Quinn, Robert Massey, and Daniel Hirsh

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.

I. Introduction

A. Background

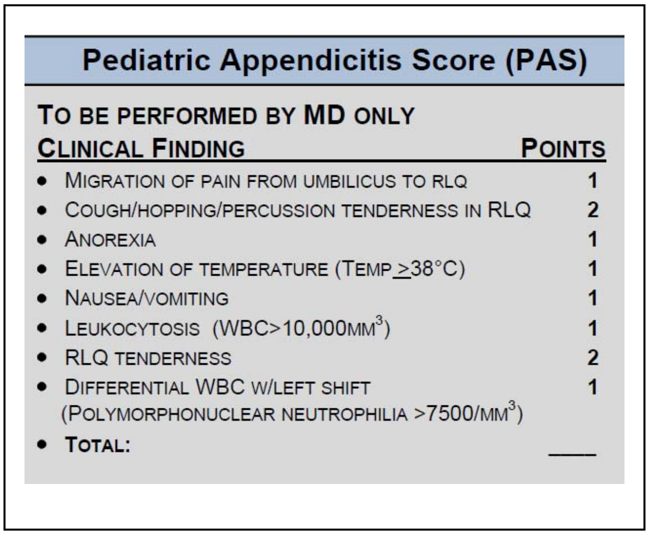
As use of electronic health records (EHR) becomes more prevalent—up from 46% of United States emergency departments in 2006 to 84% in 2011 [1]—healthcare organizations are leveraging computational methods to extract knowledge from these data. This wealth of electronic data provides numerous opportunities 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].

B. 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. 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). While the regular expression (Fig. 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. 4). Thus, the objective was to develop an NLP system capable of detecting PAS documentation in completely unstructured text.



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 $\geq$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 1. Pediatric Appendicitis Score (PAS). *Illustration of how the PAS is calculated. Total can range from 0 to 10 inclusive.*

*Project supported by Children's Healthcare of Atlanta.

B. Norman and S. Quinn are with the Institute for Artificial Intelligence, University of Georgia, Athens, GA 30602 USA (phone: 706-542-0358; fax: 706-542-8864; e-mail: bnorman@uga.edu, spq@uga.edu).

T. Davis, D. Hirsh, R. Massey, and B. Norman are with Children's Healthcare of Atlanta, Atlanta, GA 30345 USA (e-mail: Tod.Davis@choa.org, Daniel.Hirsh@choa.org, Robert.Massey@choa.org, Brittany.Norman@choa.org).

978-1-5090-4179-4/17/$31.00 ©2017 IEEE

"…**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. 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 +appendicitis +score.+?total[^[:alnum:]]*?([[:alpha:]]+|(\d+))', 1, 1, 'ni', 2)
```

Figure 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.)*

II. Methods

A. Data

This retrospective study was conducted using a collection of 15,074 electronic Emergency Department (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.

"…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 4. Free-form PAS Documentation. *Here is one example of what documentation of PAS could look like in unstructured text. The possibilities are innumerable.*

For the collection of semi-structured data (12,562 notes), the regular expression (RE) from Fig. 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.

B. 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

1) *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.

2) *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: Naive 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.

III. Results

Fig. 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.

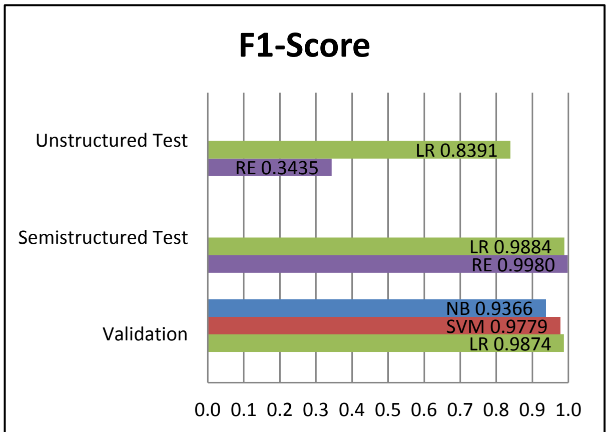
Once selected, the LR model was evaluated on the two testing sets (semistructured and unstructured) and compared with the RE approach (Fig. 5). Recall that the test sets are hand-labeled, so these scores are measured in comparison to human performance. Although the RE received the highest score (0.9980 F1) on semi-structured notes, it obtained the lowest score on unstructured notes (0.3435 F1). The LR model achieved highest performance on unstructured notes (0.8391 F1). To see how these F-scores decompose into precision and recall, see Table 1.

Table 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 to compare these features with an example of PAS documentation.

IV. 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 can be found in the example of PAS documentation from Fig. 2, thus a



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

Figure 5. F1-Score Performance. *All numerical values represent F1-scores. NB=Naïve Bayes. SVM=Support Vector Machine. LR=Logistic Regression. RE=Regular Expression*

TABLE I. 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**

NB=Naïve Bayes, SVM= Support Vector Machine, LR=Logistic Regression, RE=Regular Expression

TABLE II. 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

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, 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.

V. 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 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.

ACKNOWLEDGMENTS

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 Jessie 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.