INTERACTIVE NATURAL LANGUAGE PROCESSING FOR CLINICAL TEXT

by

Gaurav Trivedi

B. Tech, National Institute of Technology Karnataka, 2012

MS, University of Pittsburgh, 2015

Submitted to the Graduate Faculty of
the School of Computing and Information in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This dissertation was presented

by

Gaurav Trivedi

It was defended on

29th April 2019

and approved by

Dr. Harry Hochheiser, Assoc. Professor, Intelligent Systems and Biomedical Informatics

Dr. Shyam Visweswaran, Assoc. Professor, Intelligent Systems and Biomedical Informatics

Dr. Rebecca Hwa, Assoc. Professor, Intelligent Systems and Computer Science

Dr. Wendy Chapman, Professor, Department of Biomedical Informatics, University of Utah

Dissertation Director: Dr. Harry Hochheiser, Assoc. Professor, Intelligent Systems and Biomedical Informatics

INTERACTIVE NATURAL LANGUAGE PROCESSING FOR CLINICAL TEXT

Gaurav Trivedi, PhD

University of Pittsburgh, 2019

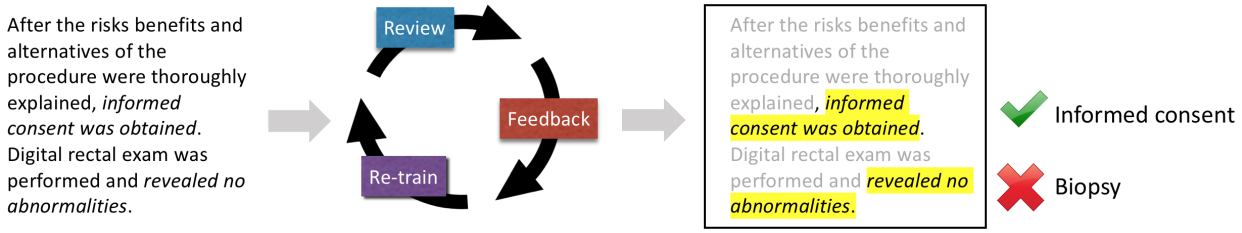

After the risks benefits and alternatives of the procedure were thoroughly explained, *informed consent was obtained*. Digital rectal exam was performed and *revealed no abnormalities*.
Review
Feedback
Re-train
After the risks benefits and alternatives of the procedure were thoroughly explained, *informed consent was obtained*. Digital rectal exam was performed and *revealed no abnormalities*.
✓ Informed consent
✗ Biopsy

Figure 1: **Interactive Natural Language Processing allows domain experts without machine learning experience to build models on their own, and also reduce or eliminate the need for collecting prior annotations and training data.**

Free-text allows clinicians to capture rich information about patients in narratives and first-person stories. Care providers are likely to continue using free-text in Electronic Medical Records (EMRs) for the foreseeable future due to convenience and utility offered. However, this complicates information extraction tasks for big-data applications. Despite advances in Natural Language Processing (NLP) techniques, building models on clinical text is often expensive and time-consuming. Current approaches require a long collaboration between clinicians and data-scientists. Clinicians provide annotations and training data, while data-scientists build the models. With the current approaches, the domain experts - clinicians and clinical researchers - do not have provisions to inspect these models or give direct feedback. This forms a barrier to NLP adoption and limits its power and utility for real-world clinical applications.

Interactive learning systems may allow clinicians without machine learning experience to build NLP models on their own. Interactive methods are particularly attractive for clinical text due to the diversity of tasks that need customized training data. Interactivity could enable end-users (clinicians) to review model outputs and provide feedback for model revisions within a closed feedback loop (Figure 1). This approach may make it feasible to extract understanding from unstructured text in patient records; classifying documents against clinical concepts, summarizing records and other sophisticated NLP tasks while reducing the need for prior annotations and training data upfront.

In my dissertation, I demonstrate this approach by building and evaluating prototype systems for both clinical care and research applications. I built NLPReViz as an interactive tool for clinicians to train and build binary NLP models on their own for retrospective review of colonoscopy procedure notes. Next, I extended this effort to design an intelligent signout tool to identify incidental findings in a clinical care setting. I followed a two-step evaluation with clinicians as study participants: a usability evaluation to demonstrate the feasibility and overall usefulness of the tool, followed by an empirical evaluation to evaluate model correctness and utility. Lessons learned from the development and evaluation of these prototypes will provide insight into the generalized design of interactive NLP systems for wider clinical applications.

TABLE OF CONTENTS

PREFACE xiv

1.0 INTRODUCTION 1

2.0 BACKGROUND AND GOALS 5

2.1 Interactive Machine Learning 5

2.2 Relationship with other types of machine learning 7

2.3 *Review*, *Feedback*, and *Retrain* 8

2.4 Interactive NLP and Clinical text 8

2.4.1 Applications in Clinical Care and Research 11

3.0 RELATED WORK 14

3.1 Interactive Machine Learning 14

3.2 Interactive Natural Language Processing 20

4.0 *NLPReViz:* INTERACTIVE NLP FOR RETROSPECTIVE REVIEW 31

4.1 Learning Model 32

4.2 Interface Design 33

4.3 Evaluation 34

4.3.1 Dataset 34

4.3.2 Participants 36

4.3.3 Protocol 36

4.4 Results 37

4.5 Discussion 38

5.0 *INTELLIGENT SIGNOUTS:* INTERACTIVE NLP IN CLINICAL CARE: IDENTIFYING INCIDENTAL FINDINGS FROM TRAUMA REPORTS 44

5.1 Signout Notes 46

5.1.1 Identifying Incidental findings for preparing signouts 50

5.1.2 Modeling as an information extraction task 52

5.2 Objectives 54

5.2.1 Design and implement a prototype interactive tool 54

5.2.2 Hypotheses 58

5.3 Methods and Materials 59

5.3.1 Dataset 59

5.3.2 Learning Pipeline 60

5.3.3 Interface Design 62

5.3.4 Implementation and Deployment 74

5.3.5 Evaluation 74

5.4 Results 76

5.4.1 Quantitative Analysis 76

5.4.2 Qualitative Analysis 78

5.5 Discussion 89

6.0 CONCLUSION 92

BIBLIOGRAPHY 97

LIST OF TABLES

1 NLPReViz (Chapter 4) and Intelligent Signouts (Chapter 5) describe two prototypes demonstrating Interactive Natural Language Processing on Clinical Text. Together they cover two different example use-cases and applications in clinical care and research. 12

2 Example design innovations to support interactive learning in prior work. They are grouped the three steps of the learning cycle: Review, Feedback, and Retrain. 16

3 Some examples of interactive systems for Natural Language Processing. They are sorted by the NLP tasks they are designed to perform. It is split into two parts: the first part describes tools that support different parts of a data scientists' workflow. The latter group consists of tools for end-users with little to no machine learning experience. 22

4 *Activity Pattern*: This table summarizes the activity patterns of the individual participants. The *p1-p4* started with an initial model trained on 10 documents, while *p5-p9* started with 30 documents. 39

5 Summary of insights for the user interface design, grouped into the three steps of the learning cycle. 42

6 *Annotation Guidelines:* Any lesion of malignant potential and any arterial aneurysm that is greater than a specified size was annotated. 61

7 Prevalence of incidental findings in the data sampled from the annotated dataset. Positives denote the raw count of sentences, sections, or reports containing one or more incidental findings. 62

8 List of prioritized solutions to address the design requirements of the signout tool. They are grouped the three steps of the learning cycle: review, feedback & retrain. 65

9 *Participants*: Description of user study participants. An average SUS (System Usability Scale) score of 78.67 was observed using post-study questionnaires. 76

10 *Final scores*: Comparison of Precision (P), Recall (R) and F1 scores between initial and final model revisions for all 15 participants. Cohen's Kappa (κ) measures the agreement between feedback provided by the participants and the gold standard labels. The initial model was trained on the same 6 patient cases to bootstrap the learning cycle. 82

11 List of design recommendations for improving the system from the user study. 88

LIST OF FIGURES

1 **Interactive Natural Language Processing allows domain experts without machine learning experience to build models on their own, and also reduce or eliminate the need for collecting prior annotations and training data.** iv

2 **Relationship between supervised, interactive machine learning, and human-in-the-loop algorithms.** 9

3 **The *Conceptual Model* illustrates a simplified view of the visual interactive labeling (VIAL) process as described by Bernard et al. Note that *'learning model'* refers to the supervised machine learning algorithm used in the system. In this dissertation, I focus on the visual interface components from the VIAL process. *Mental Model:* From a user's perspective it can be broken into three steps: 1) review the model outputs, 2) provide feedback using the labeling interface and 3) retrain to build revised models and verify changes. These steps are also indicated in the same color in the conceptual model.** 10

4 **Adopting interactive methods for different NLP problems on clinical text.** 11

5 An interactive machine learning cycle begins with the review step, with the output from the learning model displayed to the user. User feedback is used by the system to improve upon the machine learning models by providing labels for documents that were previously not part of the training set, or by correcting any misclassified documents. After retraining, a new model is created, and the tool highlights show prediction changes along with providing guidance for resolving potentially contradictory feedback items. 32

6 (a) The *Grid view* shows the extracted variables in columns and individual documents in rows, providing an overview of NLP results. Below the grid, we have statistics about the active variable with (b) the distribution of the classifications for the selected variable and (c) the list of top indicators for that variable aggregated across all the documents in the dataset. (d) Indicators from the active report are shown on the right. (e) The *document view* shows the full-text of the patient reports with the indicator terms highlighted. (f) Feedback can be sent using the yellow control bar on the top, or by using a right-click context menu. 34

7 Screenshots of the *WordTree* and *Re-Train* views. 35

8 Plots showing the variations in F1-Scores for the two variables as the participants provided feedback. These results are shown for the test dataset only. Participants p1-p4 start with an initial training set of 10 documents, while p5 onward used a model trained on 30 training documents. Differences in the spacing of the points in each graph reflect differences in feedback rates across participants. Kappa scores next to the participant ids indicate how well their feedback compared to the gold standard labels. 40

9 My new tool extends the learning task in NLPReViz (Chapter 4) from not only classifying documents to also identifying relevant text spans within them. 43

10 *System overview:* Building iterative Learning Models for predicting important and relevant information within clinical notes. Physicians 1) review highlights predicted by the system, 2) and provide feedback on them. 3) Once these feedbacks are used to re-train models, it completes an interactive learning cycle. 46

11 Example of a signout note compared with a history and physical examination (H&P) note. Signout notes often summarize information from H&P, Progress notes, Radiology reports and also Operative notes. Incidentals is one of the main sections of this note. 48

12 Example of radiology note with incidental findings highlighted. The patient was admitted into trauma after a fall from a step stool. The CT scan reveals massive volume *ascites* and *cirrhotic* changes as incidental findings. These findings are also repeated in the 'Impression' section. This is one of the 6 radiology reports for this patient. 49

13 Intelligent Signout Tool that learns from physicians' use over time. . . . 52

14 Example of a classification pipeline to identify relevant sentences used by Yetisgen et al. 57

15 1) A de-identified radiology report of CT imaging in a patient with trauma. It revealed a *nodule* as an incidental finding that is highlighted in yellow by the prototype tool (a & c). The users are able to add incidentals missed by the prototype (bolded in a) and also remove incorrectly highlighted findings (b). 2) The tool shows an overview of the patient case in a miniaturized view of all the records with highlights marking regions of interest (d). In the right sidebar, the tool allows the users to define search terms to be highlighted in pink 3) These can be seen as rules which can help attract user attention to potentially important parts of the case. 4) Shows the list of predictions made by the system. Clicking on a blurb item scrolls the report view to relevant prediction into view. 5) Shows a log of feedback items and changes recorded by the user. 72

16 (a) Users can add feedback by highlighting a span of text and triggering the contextual menu with a right click. (b) To add or remove an entire sentence, report or encounter (patient-case), the contextual menu can also be launched without manually selecting any text spans. 73

17 ***Reports:*** **Change in F1 scores over time at the report level. The colored points represent individual participants. The grey band marks the average score and tapers off in thickness to represent the number of participants completing that revision during the study.** 79

18 ***Sections:*** **Change in F1 scores over time at the section level. The colored points represent individual participants. The grey band marks the average score and tapers off in thickness to represent the number of participants completing that revision during the study.** 80

19 ***Sentences:*** **Change in F1 scores over time at the sentence level. The colored points represent individual participants. The grey band marks the average score and tapers off in thickness to represent the number of participants completing that revision during the study.** 81

20 **Average time spent in seconds between control and intervention conditions. The dots represent individual samples. We observed a statistically significant lower time in intervention vs control conditions (mean time: 134.38 vs. 148.44 seconds, Wilcoxon:** $Z = 10$, $p < 0.005$)**. One participant spent much longer time per patient case than others and can be seen as an outlier in both the conditions.** 83

21 **Average feedback counts between control and intervention conditions. The dots represent individual samples. We observed statistically significant counts in the intervention vs. control conditions (average counts: 42 vs. 55.07, Wilcoxon:** $Z = 13.5$, $p < 0.05$) 84

22 **We found no statistical differences between final F1 scores or agreement with gold standard labels between control and intervention conditions at any level.** 85

PREFACE

I would like to dedicate my dissertation to my family for their unconditional support and encouragement. This work is also dedicated to my teachers for kindling my curiosity in science and engineering.

Work discussed in this dissertation is a result of collaborative efforts between many groups. I am grateful to the Intelligent Systems Program for supporting multidisciplinary Artificial Intelligence research to encourage such collaborations, especially at the Department of Biomedical Informatics and the University of Pittsburgh Medical Center. I would like to acknowledge the support of my advisor and my co-authors. I also thank Intelligent System Program, Department of Biomedical Informatics, and Office of the Provost at University of Pittsburgh for supporting my research through fellowships.

1.0 INTRODUCTION

Electronic Medical Record (EMR) systems should emphasize ease of use and utility for clinicians for over considerations for data analysis. Free-text makes it convenient for recording patient conditions and progress in narratives and first-person stories [1, 2]. As a result, literature in clinical informatics suggests that a large portion of information in the EMRs, which is relevant to research and clinical outcomes, is locked in free-text [3, 4]. The American Medical Association's call for design overhaul of Electronic Medical Records also includes reducing cognitive burden due to information overload in their top eight priorities [5]. Within EMRs, prior research identifies clinical notes in the electronic medical record [6, 7, 8] as one of the culprits for this problem. While narratives and free-text allow physicians to easily capture rich information [1, 2], they are difficult to analyze. Solving these problems become even more important as we increasingly depend on a team of care providers responsible for a patient's well-being. Patient records are collaboratively created and reviewed by large teams of physicians and nurses. These records can become bloated quickly and make it harder for care providers to find relevant pieces of information [8]. The current documentation processes may be responsible for producing more bloated records than what one would otherwise need for each individual use case [9]. Wier and Nebeker [6] point out that there are multiple uses cases for the same component of the record:

1. They are the main source of communication regarding the overall *plan of care* between providers.
2. They are also used by *coders for billing* purposes.
3. Plus, they form the main source of information for *legal and quality review*.

The large and increasing volume of documentation results in a higher workload and also potential sources of error [10]. As a part of their study, Weir and Nebekar interviewed 88 providers across different settings about the documentation practices. Their results identified difficulties in finding information and extracting meaningful data as significant problems:

> “[clinicians] commented on the difficulty of sorting, sifting and locating relevant documents.”

> “Most providers reported on the difficulty of extracting meaningful data from a large number of notes and confusing text.”

> “...they expect relevant information to be displayed and non-relevant information to be omitted.”

These results are reiterated in more recent studies, such as by Wright et al. [7]:

> “custom-filtered displays, intended to provide quick access to frequently used information for specific clinicians, fell short of their needs.”

And, also by Artis et al. [11], while discussing the use of the patient records for activities such as preparing sign-out notes and conducting daily rounds:

> “... [EHR system] does not automatically provide an effective visual display of data needed for daily rounds”

NLP could help alleviate the problem of information load by identifying relevant and important information while omitting other pieces depending on user needs. Using Natural Language Processing (NLP) on clinical text require collaboration between clinicians (domain experts) and data scientists (NLP experts). Clinicians who can understand medical jargon and create training examples are usually not versed in informatics techniques to be able to use NLP directly. Clinicians provide training data, while data scientists build and evaluate appropriate models. As a result, despite recent advances in Natural Language Processing (NLP) techniques, building models is often expensive and time-consuming [12] as it requires expert construction of gold standard and training corpora [13]. These steps involved in building NLP models do not generalize and must be repeated for every specific task or application. Current tools also lack provisions for domain experts to inspect NLP outcomes and make corrections that might improve these results. Due to these factors, Chapman et al. [14] have identified "lack of user-centered development as one of the barriers to NLP adoption in the clinical domain.

EMRs have long been recognized as vital sources of information for decision support systems, data-driven quality measures, and many other applications [15]. An intelligent EMR system could learn from the users' usage patterns and build predictive models for identifying relevant pieces of data. This could not only improve clinical care but also further clinical research faster. However, the extraction of information from unstructured clinical notes presents many challenges due to their predominantly free-text nature [16]. The biggest obstacle in building such systems that can learn using EMR data is the lack of labeled training data [17]. There have been few efforts towards exploring "human-in-the-loop" and interactive methods which reduce the need for labeled examples upfront and bring machine learning closer to clinician end-users.

In this dissertation, I seek to evaluate whether interactive tools can bridge this gap and make NLP more valuable for applications in both clinical care and research. To this end, I design, implement and evaluate prototype tools for users with little or no machine learning experience. I first demonstrate this approach with an example application in clinical research. Building on this work, I then explore an example use-case of interactive NLP in clinical care. Insights from building these prototypes can help us generalize our approach for a wider range of NLP problems on clinical text. This would help us move closer towards unlocking the full potential of analyzing free-text notes in the long run.

The following chapter (Chapter 2) lays out the goals of my dissertation as well as the background work that influence my work. It lays out how interactive machine learning can be used to address the barriers to NLP on clinical notes. In Chapter 3, I present a detailed literature review of related work. I describe my work demonstrating the interactive NLP approach for retrospective research on colonoscopy procedure notes in Chapter 4. It presents NLPReViz: an interactive web-based tool to allow clinicians and clinical researchers to build NLP models for binary concepts at the document level. Building upon this work, I prototype an intelligent signout tool that can identify relevant portions of clinical text in a patient's electronic record (Chapter 5). This work further refines the NLP task from making document level predictions to identifying relevant text spans within a document and illustrates the interactive approach in a clinical care setting. I conclude this dissertation in Chapter 6 with a discussion on the two prototype tools, their limitations and directions for future work.

2.0 BACKGROUND AND GOALS

2.1 INTERACTIVE MACHINE LEARNING

Traditionally machine-learning is classified into *supervised* and *unsupervised* learning families. In supervised learning, training data, $\mathcal{D}$, consists of N sets of feature vectors – each with a desired label:

Training set $\quad \mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$

where, $\mathbf{x}_i \in \mathcal{X}$ is a d-dimensional feature vector and $y_i \in \mathcal{Y}$ is the known label. The task is to learn a function, $f : \mathcal{X} \rightarrow \mathcal{Y}$, which can be used on unseen data.

In unsupervised learning, the data consists of the vectors $\mathbf{x}_i$ but not the target label y_i. Common tasks under this category include clustering, density estimation, and pattern discovery. A combination of these two classes is called *semi-supervised* learning, which has a mixture of labeled and unlabeled data in the training set. The algorithm assigns labels for missing data points using certain similarity measures.

Supervised machine learning has been the dominant form of learning. However, traditional supervised algorithms assume that the training data along with their corresponding labels are readily available. They are not concerned about the process of obtaining the target labels y_is in the training dataset. Often, obtaining labeled data is one of the main bottlenecks in applying these techniques in domain specific applications. Further, current approaches do not provide easy mechanisms for the end-users to correct problems with the models. In Natural Language Processing, models are often built by experts in linguistics and/or machine learning, with limited or no scope for the end-users to provide input. Here the domain experts, or the end-users, provide input to models as annotations for a large batch of training data. This approach can be expensive, inefficient and even infeasible in many situations [17]. This includes many problems in the clinical domain such as building models for analyzing EMR data.

"Human-in-the-loop" methods incorporate human input to guide the learning algorithms [18]. Such methods be able to leverage capabilities of the domain experts through user interaction [19]. Interactive Machine Learning (IML) is a subset of this class of algorithms. It is defined as the process of building machine learning models iteratively through end-user input. Some other IML definitions found in the literature in increasing order of refinement are as follows:

"...interactive machine learning engages users [domain expert] *in actually generating the classifier themselves..."* [20]

"...interactive machine-learning (IML) model that allows users to train, classify/view and correct the classifications..." [21]

"...algorithms that can interact with both computational agents and human agents and can optimize their learning behavior through these interactions." [19]

"...a process that involves a tight interaction loop between a human and a machine learner, where the learner iteratively takes input from the human, promptly incorporates that input and then provides the human with output impacted by the results of the iteration..." [22]

A popular example of interactive learning is its application in teaching email clients about spam vs. important email. Other examples found in the literature include bug triaging [23], tailoring music and movie recommender systems [24] and even music composition [25]. Some applications may allow users to passively teach the machine learning system while they perform other tasks. On the other end of the spectrum, we have tools for users vested solely in the task of building machine learning models. Such interactive machine learning tools allow the users to review model outputs and make corrections by giving feedback for building revised models. The users are then able to see and verify model changes. Some early examples for this definition include applications in image segmentation [21], interactive document clustering [26], and document retrieval [27].

2.2 RELATIONSHIP WITH OTHER TYPES OF MACHINE LEARNING

Interactive machine learning builds on a variety of different styles of learning algorithms:

1. *Active Learning:* Active learning algorithms optimize for lowering the number of training labels y_i. Active learning-based systems ask 'oracles' (or human-experts) to provide labels such that it can achieve higher accuracy with minimum number of queries [28]. This is useful in cases where collecting training labels is expensive and time-consuming. I consider active learning to be a subset of interactive learning. While active learning primarily focuses on asking users queries about what the model needs, interactive learning is the user-centered approach for it.
2. *Reinforcement Learning:* In this class of learning, we still want to learn $f : \mathcal{X} \rightarrow \mathcal{Y}$ but we see samples of $\mathbf{x}_i$ but no target output y_i. Instead of y_i, we get a feedback from a critic about the "goodness" (reward) of the predicted output. The goal of the learner is to optimize for the reward function by selecting outputs that get best scores from the critics [29]. The critic can be a human or any other agent. There need not be a human-in-the-loop for a reinforcement learning algorithm.
3. *Online Algorithms:* Online learning (or sequential learning) algorithms are used when training data is available in sequential order, say due to the nature of the problem or memory constraints, as opposed to a batch learning technique where all of the training data is available at once. The algorithm must adapt to the continuous stream of data made available to it. Formulating the learning problem to handle this situation forms the core of algorithm design under this class. A commonly found example would be the online gradient descent method for linear regression [30]. This can be summarized as follows: Suppose we want to learn the parameters $\mathbf{w}$ for $f(\mathbf{x}) = w_0 + w_1x_1 + \ldots w_dx_d$. We update the weights when we receive the ith training example by taking the gradient of the defined error function, with α as the learning rate: $\mathbf{w}_{new} \leftarrow \mathbf{w} - \alpha \times \Delta_{\mathbf{w}} Error_i(\mathbf{w})$.

An Interactive machine learning system can include all or some of these learning techniques. Figure 2 shows a Venn diagram representing my understanding of the relationship between them. The common property between all the interactive machine learning methods is the tight interaction loop between the human and the learning algorithm.

2.3 *REVIEW*, *FEEDBACK*, AND *RETRAIN*

Jü et al. [31] detail a unified process for visual interactive labeling for model building (Figure 3). From a user's perspective it can be broken into three steps (see *Mental Model* in Figure 3):

1. ***review*** the model outputs,
2. provide ***feedback*** using the labeling interface, and
3. ***retrain*** to build revised models and verify changes.

This feedback loop allows them to refine models with every iteration. Thus, Interactive machine learning systems require effective displays for presenting outputs, eliciting user feedback, and showing model revisions. Interactive learning systems use overviews, filters, navigation support, and other information visualization techniques extensively to support these steps [32, 33, 34, 21]. These are described in more details in Chapter 3.

From the prior work in interactive machine learning, we observe that designing interfaces of each step of the interaction loop forms the critical part of the system development. We need novel user interfaces and interaction design to address these challenges. Specifically, we need to support the review, feedback and the retrain steps of the interactive learning cycle (Figure 3).

2.4 INTERACTIVE NLP AND CLINICAL TEXT

Interactive methods are particularly appealing in addressing the challenges inherent in developing NLP applications, which are further exacerbated by differences across institutions and clinical sub-domains. In the traditional approach, models are built by NLP experts in linguistics and machine learning while subject matter domain experts (clinicians, lawyers,

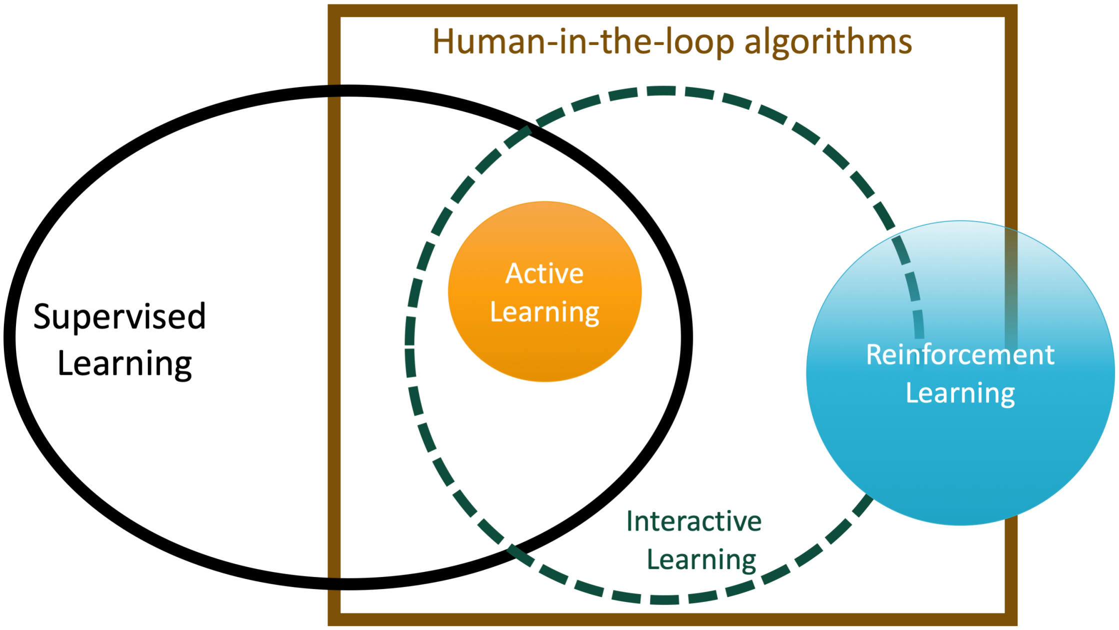


Figure 2: **Relationship between supervised, interactive machine learning, and human-in-the-loop algorithms.**

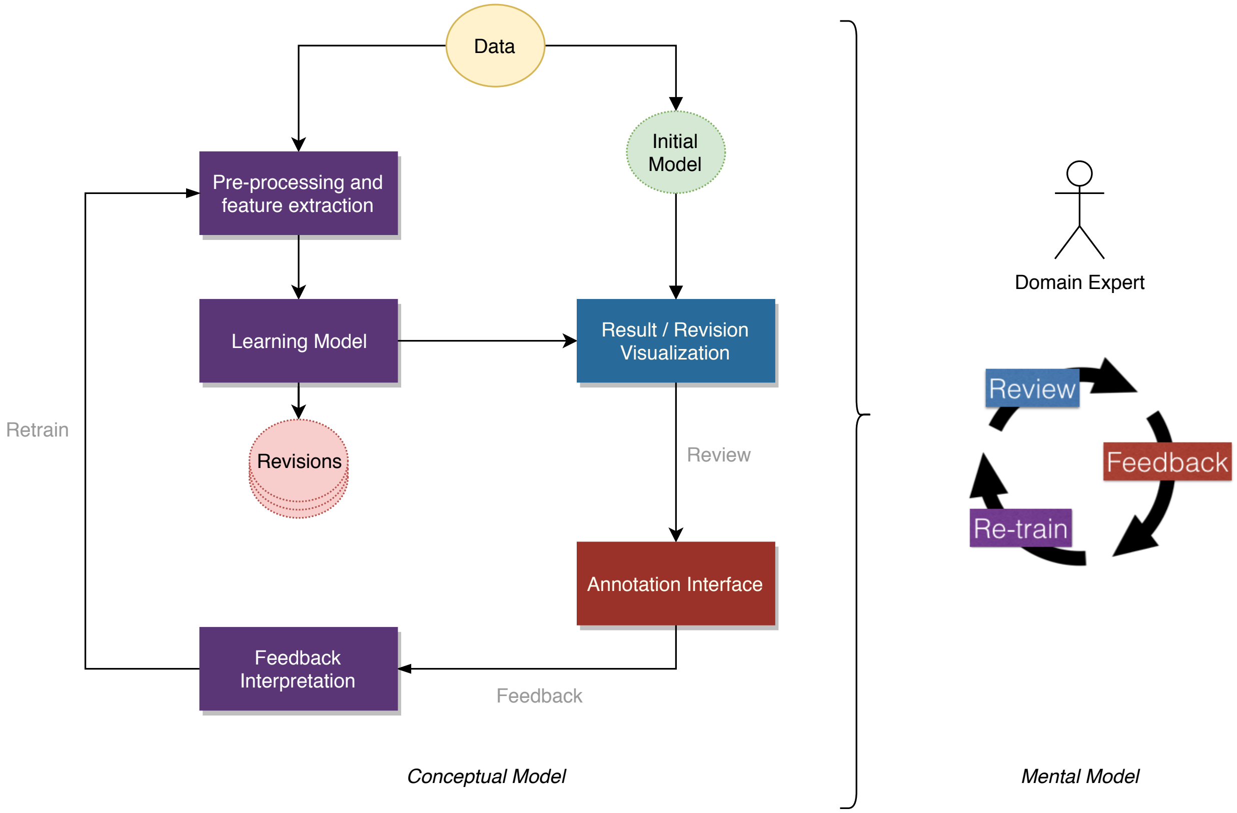


Figure 3: **The *Conceptual Model* illustrates a simplified view of the visual interactive labeling (VIAL) process as described Bernard et al.** [31]**. Note that *'learning model'* refers to the supervised machine learning algorithm used in the system. In this dissertation, I focus on the visual interface components from the VIAL process. *Mental Model:* From a user's perspective it can be broken into three steps: 1) review the model outputs, 2) provide feedback using the labeling interface and 3) retrain to build revised models and verify changes. These steps are also indicated in the same color in the conceptual model.**

etc.) who are often the end-users must construct training data through laborious annotation of sample texts. This approach is expensive and inefficient, particularly when language subtleties necessitate multiple iterations through the annotation cycle (as is often the case). For clinical applications, it quickly becomes infeasible to customize models for every specific task and application. Thus, we need user-centered tools that can address the needs of clinicians and clinical researchers, in order to make NLP more useful on clinical text. Interactive NLP tools that provide end-users with the ability to easily label data, refine models and review the results of those changes have the potential to lower the costs associated with the customization, and therefore to increase the value of NLP on clinical reports. There is scope to build interactive NLP for tasks such as classification, information extraction, summarization, named-entity recognition and relationships, question-answering, and so on (Figure 4).

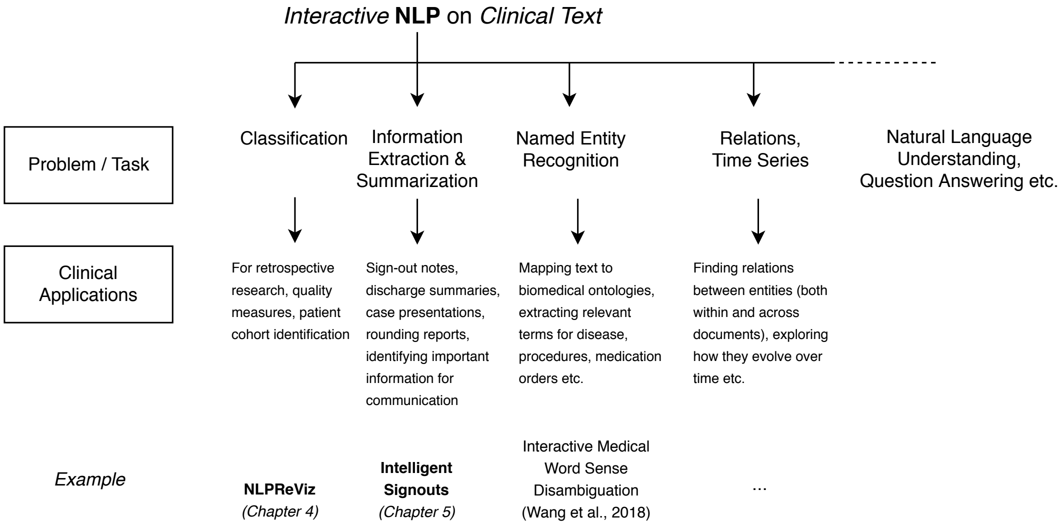

Interactive **NLP** on *Clinical Text*
Problem / Task
Clinical Applications
Example
Classification
Information Extraction & Summarization
Named Entity Recognition
Relations, Time Series
Natural Language Understanding, Question Answering etc.
For retrospective research, quality measures, patient cohort identification
Sign-out notes, discharge summaries, case presentations, rounding reports, identifying important information for communication
Mapping text to biomedical ontologies, extracting relevant terms for disease, procedures, medication orders etc.
Finding relations between entities (both within and across documents), exploring how they evolve over time etc.
NLPReViz
(Chapter 4)
Intelligent Signouts
(Chapter 5)
Interactive Medical Word Sense Disambiguation (Wang et al., 2018)
...

Figure 4: **Adopting interactive methods for different NLP problems on clinical text.**

2.4.1 Applications in Clinical Care and Research

Biomedical activities may be categorized into research or clinical care/practice [35, 36]. Interactive Natural Language Processing has promise in both improving clinical care and further clinical research faster. My dissertation explores the use of end-to-end interactive

Table 1: **NLPReViz (Chapter 4) and Intelligent Signouts (Chapter 5) describe two prototypes demonstrating Interactive Natural Language Processing on Clinical Text. Together they cover two different example use-cases and applications in clinical care and research.**

	NLPReViz	Intelligent Signout
Scenario	Retrospective research	Clinical care
End-Users	Clinical researchers vested solely in the task of building models	Clinicians engaged in patient care who are building NLP models as a background task, in addition to their primary task (eg. creating signouts)
Specific Task	Binary classification of documents	Binary classification of sentences to highlight text spans that make a document relevant
Application	Review of colonoscopy procedure notes for extracting quality metrics	Sign-out note preparation summarizing important information
Variables	Different model for each concept (metric) such as *biopsy*, *informed-consent*, etc.	Models to identify sentences for different sections of the signout note.

NLP for clinical applications and demonstrates its uses with an example application each in clinical research and care.

2.4.1.1 Clinical Research In Chapter 4, I demonstrate interactive NLP through the task of document classification in my work on NLPReViz for classifying colonoscopy reports. NLPReviz is a prototype tool for helping clinicians build models for binary classification of clinical documents. It implements the interactive learning cycle described above (*review*, *feedback*, and *retrain*). It serves as an example of how clinicians, without any prior machine learning experience, could train their own models for retrospective research.

2.4.1.2 Clinical Care In Chapter 5 presents an interactive system for identifying sentences or phrases within a note. This task is particularly relevant for the construction of summaries, which are often manually curated by clinicians for a variety of clinical tasks. Examples include writing discharge summaries, preparing for rounding, and seeing new consults. These manually curated summaries help clinicians better manage a patient's growing record. By building tools that integrate NLP, and more generally machine learning, into clinical workflows, we can address the problem of lack of upfront labeled training data and providing end-users with the ability to customize models. Interactive approaches also support the evolution of guidelines and associated models over time.

Table 1 presents a comparison between NLPReViz (retrospective review of colonoscopy procedure notes for quality metrics) and the new prototype (intelligent signout tool). They serve as example tasks where interactive NLP can be adopted for analyzing clinical text, together covering both clinical care and research applications. Together they cover a larger space for interactive NLP applications on clinical text. My goal is to use these demonstrations to understand the design of interactive NLP tools for users with no machine learning experience. This would help us generalize our approach for a wider range of analytics problems on clinical text and help address the barriers to NLP in the clinical domain. By building interactive NLP tools that focus on clinicians as end-users, we are able to more fully realize the true potential of using NLP for real-world clinical applications.

3.0 RELATED WORK

3.1 INTERACTIVE MACHINE LEARNING

In their papers, Amershi et al. [22, 34] and Kulesza et al. [37] have discussed the approaches used for interactively obtaining annotations from the users in great detail. Together they describe the design techniques in interactive machine learning and how they differ from the traditional systems. The traditional supervised learning approaches often hide the complexities of machine learning algorithms from its end-users. Amershi et al. have used prior work as case studies and focused on how human annotators can be asked to provide inputs. They describe how IML systems find motivation in the need for "developing novel feedback techniques for interactively incorporating domain expert knowledge" in model building. Prior work in IML focuses on the study of end-users to build better user interfaces to support model building. The paper also talks about the desired and undesired characteristics for designing user interactions within IML systems. Further, it describes novel interfaces or IML that can leverage domain experts' inputs in an efficient manner. These include supporting better data selection, such as in CueFlik [38], which sorts the unlabeled examples according to best and worst prediction scores for one class to help users better train the systems. Other ideas involve variations of active learning where the system intermittently queries the users about the learning problem [39], novel ways to solicit user feedback [40], allowing users to decide trade-offs – such as between precision and recall in a classification problem [41]. Table 2 describes these prior works and summarizes the design innovations to support the different steps of the interactive learning loop. Both [22] and [42] provide a good survey of prior work done in Interactive Machine Learning. Relevant to my dissertation work, I have summarized the discussion in these papers as the following two research questions:

1. *How can interfaces make the annotation process more efficient?*

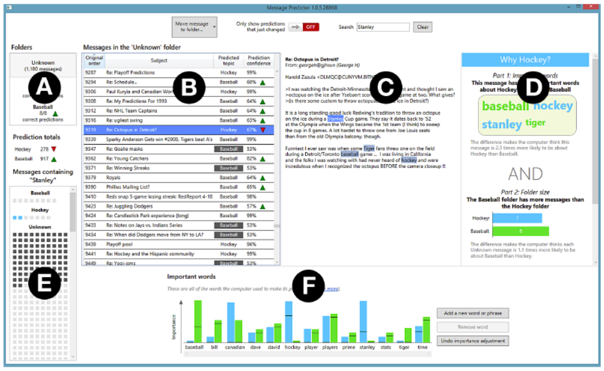
   Human annotators prefer giving rich feedback instead of merely acting as passive oracles. Simple active learning approaches involve asking a series of questions to human oracles and can thus be annoying and frustrating for them [39]. An interactive learning system should not only try to maximize the information gain from new labels but also allow users to illustrate the concept to be learned as efficiently as possible. In order to achieve this goal, we need to consider not only what inputs the users are capable of providing to the system but also how they'd be willing to do so. For example, Amershi et al. also point out that human annotators prefer more natural ways of teaching instead of providing simple 'yes-no' labels [43, 44]. Current practices can be improved by designing human-centered methods of providing feedback. These should be implemented to optimize for efficiency in terms of required time and effort from the users. Amershi et al. remark that the users will only provide feedback when they perceive the benefits of producing model revisions outweigh the costs involved.

2. *How can interactive tools help end-users understand the impact of their feedback towards building models?*

   A motivating goal for interactive systems is to allow domain experts lacking experience in machine learning to build models on their own. Often making 'black box' systems explainable helps the annotators provide better feedback using visual interface. This is demonstrated by Kulesza et al. [37] in their work on personalizing interactive machine learning for NLP models, through a combination of corpus overviews and *explanatory debugging* tools capable of explaining reasons for predictions to the end user. This enables users to offer better labels to help improve the performance of the system [45]. Another design pattern useful for interactive learning in prior work is to treat modeling as an iterative process. This allows users to work with models and their revisions. Many of the systems also offer ideas to help the users understand changes between revisions and allow rollbacks.

In Table 2, I summarize innovations described in prior work in interactive machine learning to address these requirements. These are broken down into the individual steps of the interactive machine learning cycle. I group these innovations by the three steps, as not all of the systems described in the prior work devote equal attention to them. In the following sections, I will narrow down further on the work done in interactive natural language processing leading towards the identification of gaps and opportunities that motivate my work.

Table 2: **Example design innovations to support interactive learning in prior work. They are grouped the three steps of the learning cycle: Review, Feedback, and Retrain.**

Step	Task / Paper	Description
Review	Text Classification	Supports end-user debugging by allowing users to visualize feature weights in a Naive Bayes model.
	EluciDebug [37, 46]	

Original order	Subject	Predicted topic	Prediction confidence
9287	Re: Playoff Predictions	Hockey	99%
9294	Re: Schedule..	Baseball	60% ▲
9306	Paul Kuryia and Canadian Wor	Hockey	99%
9308	Re: My Predictions For 1993	Baseball	64% ▲
9312	Re: NHL Team Captains	Baseball	64% ▲
9316	Re: ugliest swing	Baseball	63% ▲
9319	Re: Octopus in Detroit?	Hockey	67% ▼
9339	Sparky Anderson Gets win #2000, Tigers beat A's	Baseball	99%
9347	Re: Goalie masks	Baseball	53%
9362	Re: Young Catchers	Baseball	82% ▲
9371	Re: Winning Streaks	Baseball	53%
9379	Royals	Baseball	64% ▲
9390	Phillies Mailing List?	Baseball	65% ▲
9410	Reds snap 5-game losing streak: RedReport 4-18	Baseball	98%
9423	Re: Juggling Dodgers	Baseball	57% ▲
9424	Re: Candlestick Park experience (long)	Baseball	99%
9433	Re: Notes on Jays vs. Indians Series	Baseball	53%
9434	Re: When did Dodgers move from NY to LA?	Baseball	53%
9439	Playoff pool	Hockey	96%
9441	Re: Hockey and the Hispanic community	Hockey	99%
9449	Re: Yogi-isms	Baseball	53%

Re: Octopus in Detroit?
From: *georgeh@gjhsun (George H)*

Harold Zazula <DLMQC@CUNYVM.BITN

>I was watching the Detroit-Minnesota nt and thought I saw an
>octopus on the ice after Ysebaert scor game at two. What gives?
>(is there some custom to throw octopus ice in Detroit?)

It is a long standing good luck Redwing's tradition to throw an octopus on the ice during a Stanley Cup game. They say it dates back to '52 at the Olympia when the Wings became the 1st team (I think) to sweep the cup in 8 games. A lot harder to throw one from Joe Louis seats than from the old Olympia balcony, though.

Funniest I ever saw was when some Tiger fans threw one on the field during a Detroit/Toronto baseball game ... I was living in California and the folks I was watching with had never heard of hockey and were incredulous when I recognized the octopus BEFORE the camera closeup !!

Part 1: Im words
This message has tant words about Hockey Baseball

baseball hockey stanley tiger

The difference makes the computer think this message is 2.3 times more likely to be about Hockey than Baseball.

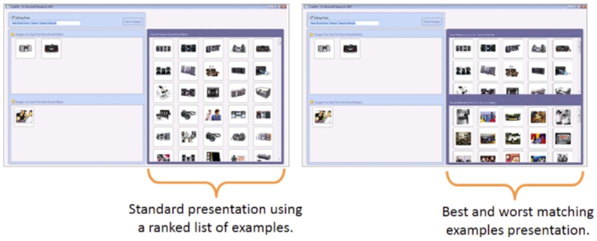
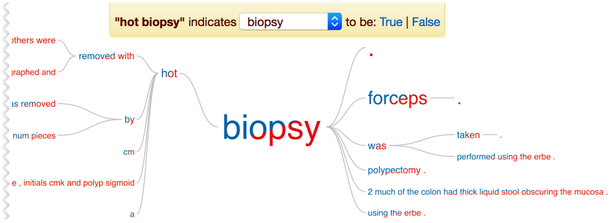
AND

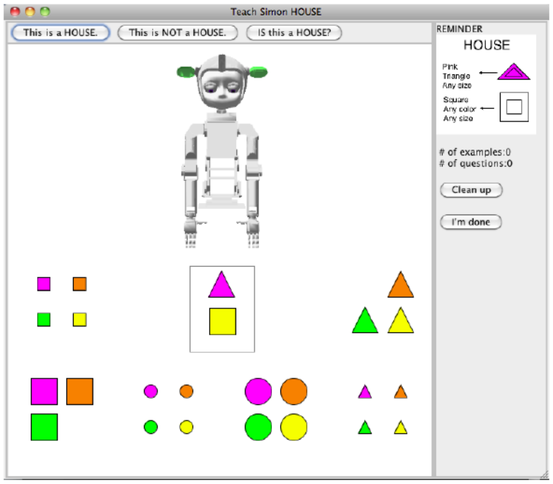
Part 2: Folder size
The Baseball folder has more messages than the Hockey folder

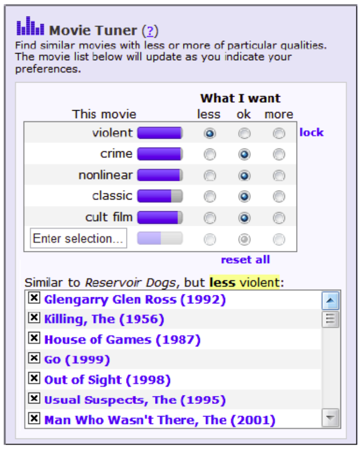
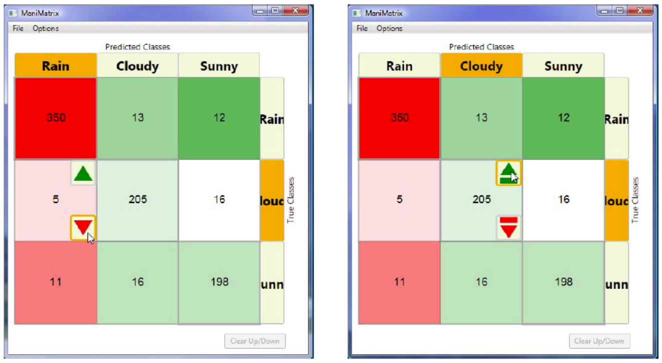
Hockey:	7
Baseball:	8

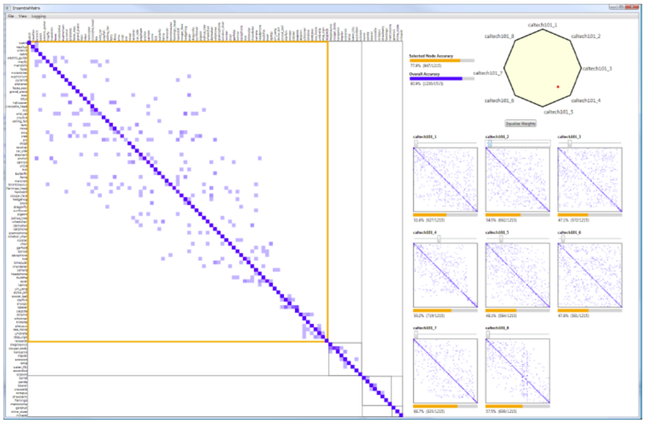
The difference makes the computer thinks each Unknown message is 1.1 times more likely to be about Baseball than Hockey.

These are all of the words the computer used to make its p more)

	Image Retrieval (Binary Classification) CueFlik [38] Overview Based Selection [47]	Show only the best and the worst matching examples with the current model. This enabled users to train better than presenting with all of the data showing a ranked list of examples.  Standard presentation using a ranked list of examples. Best and worst matching examples presentation. [47] further extends this to show an overview representation of multiple image clusters. They are grouped based on similarity in image concepts they represent.
Feedback	Text Classification NLPReViz [48]	Designed three different kinds of feedback mechanisms: a) assign labels for the whole document, b) highlight text spans as evidence or rationales, and c) using the WordTree, which is useful for providing feedback on several documents together in an efficient manner.  "hot biopsy" indicates biopsy to be: True | False others were removed with raphed and s removed by num pieces cm e , initials cmk and polyp sigmoid a hot biopsy . forceps . was taken . performed using the erbe . polypectomy . 2 much of the colon had thick liquid stool obscuring the mucosa . using the erbe .

	Shape Classification (application in Human-Robot Interaction) Teaching Simon [39]	Cakmak et al. observed that when the inputs were initiated by users, it resulted in learning as fast as the fully active learning approach. Fully active learning approach involved the system asking a continuous stream of questions to the user. Interactions triggered by the users followed by intermittent active-learning style questions were rated better by human teachers than continuous queries.  Teach Simon HOUSE This is a HOUSE. This is NOT a HOUSE. IS this a HOUSE? REMINDER HOUSE Pink Triangle Any size Square Any color Any size # of examples:0 # of questions:0 Clean up I'm done

	Movie Recommender System Movie Tuner [24]	Users may specify levels of emphasis on specific tags (similar to features in a machine learning formulation). For example, users could specify whether they wanted less *violent* and more of a *cult film*, where *violent* and *cult film* are tags describing the movies.  **Movie Tuner** (?) Find similar movies with less or more of particular qualities. The movie list below will update as you indicate your preferences. **What I want** This movie less ok more violent ● ○ ○ **lock** crime ○ ● ○ nonlinear ○ ● ○ classic ○ ● ○ cult film ○ ● ○ Enter selection... ○ ● ○ **reset all** Similar to *Reservoir Dogs*, but **less** violent: ☒ **Glengarry Glen Ross (1992)** ☒ **Killing, The (1956)** ☒ **House of Games (1987)** ☒ **Go (1999)** ☒ **Out of Sight (1998)** ☒ **Usual Suspects, The (1995)** ☒ **Man Who Wasn't There, The (2001)**
Retrain	Multi-class Classification ManiMatrix [41]	Allows users to decide trade-offs – such as between precision and recall in a classification problem. Users can increase or decrease the different types of errors using the interface.  ManiMatrix File Options Predicted Classes Rain Cloudy Sunny 350 13 12 Rain 5 205 16 louc 11 16 198 unn True Classes Clear Up/Down ManiMatrix File Options Predicted Classes Rain Cloudy Sunny 350 13 12 Rain 5 205 16 louc 11 16 198 unn True Classes Clear Up/Down

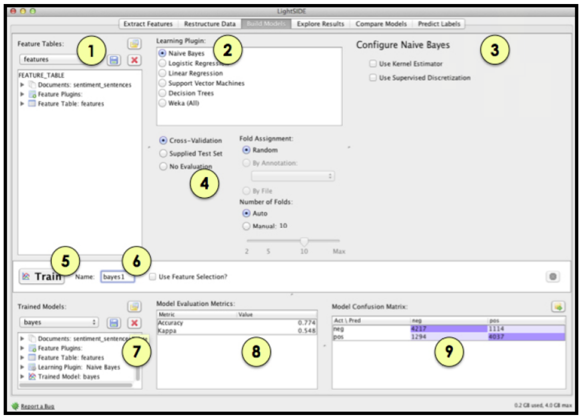
	Ensemble Learning Ensemble Matrix [49]	Visualizes the confusion matrices to help users understand the relative merits of individual learners. Users can then specify weights in the linear combination of learners to build an ensemble model.  Selected Node Accuracy Overall Accuracy caltech101_1 caltech101_2 caltech101_3 caltech101_4 caltech101_5 caltech101_6 caltech101_7 caltech101_8 Equalize Weights
	Text Classification NLPReViz [48]	The interface lists any potential inconsistencies and conflicts in the user-provided feedback. After retraining, it highlights documents with changes in variable assignments due to model revisions.

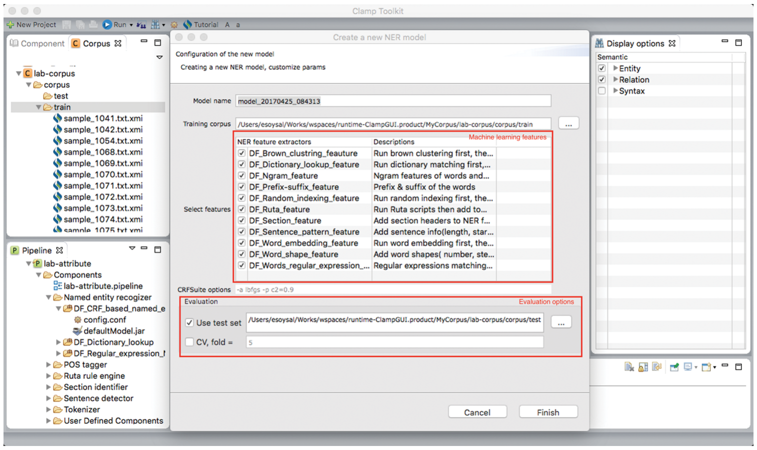
3.2 INTERACTIVE NATURAL LANGUAGE PROCESSING

Table 3 provides a summary of tools for building NLP models interactively. It includes applications across different domains. Within the biomedical informatics domain, there have been efforts to make machine learning on clinical notes interactive. Some examples include work done on interactive medical word sense disambiguation [50] to allow users to specify indicative words of a sense and highlight supporting evidence. The paper demonstrated that the interactive approach outperforms traditional labels using only active learning and requires much less labeling effort. Similarly, RapTAT demonstrated how interactive annotation can be used to reduce the time required to create an annotated corpus by learning to pre-annotate documents [51, 52]. Other efforts are towards helping users define the parameters and configurations of the different components of the NLP pipeline. D'Avolio et al. [53] describe a prototype system called ARC that combines several existing tools such for creating text annotations (Knowtator [54]), and for deriving NLP features (cTAKES [55]), using a common user interface that can be used to configure the machine learning algorithms and export their results. More recently, CLAMP toolkit [56] was developed keeping this goal in mind, for example. Most of these tools provide graphical user interfaces (GUIs) to be used by data scientists or NLP experts but do not address the challenges in designing end-to-end interactive systems with clinician end-users. My dissertation work complements these efforts by focusing not only customizing individual components of the NLP pipeline but addressing the design of different components required for building closed loop interactive machine learning systems for clinical text.

I have split Table 3 into two parts: the first part describes tools that support different parts of a data scientists' workflow. The latter group consists of tools for end-users with little to no machine learning experience. For this group, I also present an explanation for their coverage of individual steps in the cycle.

Table 3: **Some examples of interactive systems for Natural Language Processing. They are sorted by the NLP tasks they are designed to perform. It is split into two parts: the first part describes tools that support different parts of a data scientists' workflow. The latter group consists of tools for end-users with little to no machine learning experience.**

1. *Toolkits* for use by data-scientists	
Task / Publication	**Description**
Toolkit LightSIDE [57]	GUI support for machine-learning and feature-extraction similar to Weka [58], but specifically designed for text classification pipelines. Intended to be used as a researchers' workbench. Provides support for selecting algorithms, tuning parameters and viewing evaluation metrics.  LightSIDE Extract Features | Restructure Data | Build Models | Explore Results | Compare Models | Predict Labels Feature Tables: features FEATURE_TABLE ▸ Documents: sentiment_sentences ▸ Feature Plugins: ▸ Feature Table: features Learning Plugin: ● Naive Bayes ○ Logistic Regress ○ Linear Regression ○ Support Vector Machines ○ Decision Trees ○ Weka (All) Configure Naive Bayes □ Use Kernel Estimator □ Use Supervised Discretization ● Cross-Validation ○ Supplied Test Set ○ No Evaluation Fold Assignment: ● Random ○ By Annotation: ○ By File Number of Folds: ● Auto ○ Manual: 10 2 5 10 Max Train Name: bayes1 □ Use Feature Selection? Trained Models: bayes ▸ Documents: sentiment_sentence ▸ Feature Plugins: ▸ Feature Table: features ▸ Learning Plugin: Naive Bayes ▸ Trained Model: bayes Model Evaluation Metrics: Metric | Value Accuracy | 0.774 Kappa | 0.548 Model Confusion Matrix: Act \ Pred | neg | pos neg | 4217 | 1114 pos | 1294 | 4037 Report a Bug 0.2 GB used, 4.0 GB max

Toolkit CLAMP [56]	Graphical interface for building customized NLP pipeline including annotation, modeling, and processing. CLAMP's components can tackle several commonly used NLP tasks such as: a) sentence-boundary detection, tokenization, part-of-speech tagger, section header identification, word-sense disambiguation, and others. These components can be selected and added to an NLP pipeline using a drag and drop interface. It provides an annotation interface for specific modules such as Named-Entity Recognition for building CRF models. 

Create a new NER model

Configuration of the new model

Creating a new NER model, customize params

Model name: model_20170425_084313

Training corpus: /Users/esoysal/Works/wspaces/runtime-ClampGUI.product/MyCorpus/lab-corpus/corpus/train

Select features

	NER feature extractors	Descriptions
☑	DF_Brown_clustring_feauture	Run brown clustering first, the...
☑	DF_Dictionary_lookup_feature	Run dictionary matching first,...
☑	DF_Ngram_feature	Ngram features of words and...
☑	DF_Prefix-suffix_feature	Prefix & suffix of the words
☑	DF_Random_indexing_feature	Run random indexing first, the...
☑	DF_Ruta_feature	Run Ruta scripts then add to...
☑	DF_Section_feature	Add section headers to NER f...
☑	DF_Sentence_pattern_feature	Add sentence info(length, star...
☑	DF_Word_embedding_feature	Run word embedding first, the...
☑	DF_Word_shape_feature	Add word shapes(number, ste...
☑	DF_Words_regular_expression_...	Regular expressions matching...

CRFSuite options: -a lbfgs -p c2=0.9

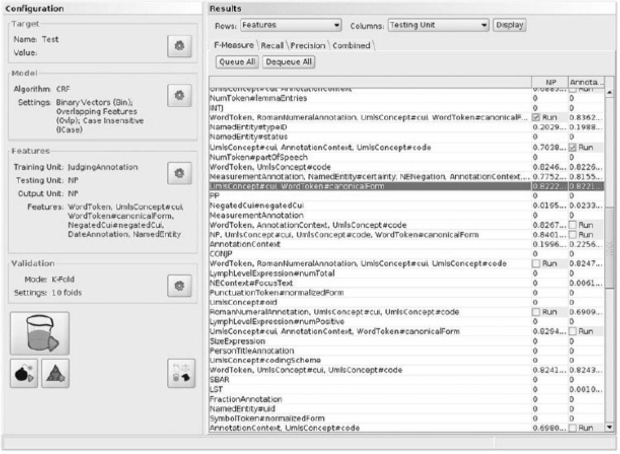
Evaluation

☑ Use test set: /Users/esoysal/Works/wspaces/runtime-ClampGUI.product/MyCorpus/lab-corpus/corpus/test

□ CV, fold = 5

<table>
<tr>
<td><em>Information extraction & Named Entity Recognition</em>

Automated Retrieval Console (ARC) [53]</td>
<td>ARC combines several existing tools such as Knowtator [54] for creating text annotations and cTAKES [55] for deriving NLP features. This common user interface also provides support for the users to configure parameters of machine learning algorithms for rapid development of NLP models.

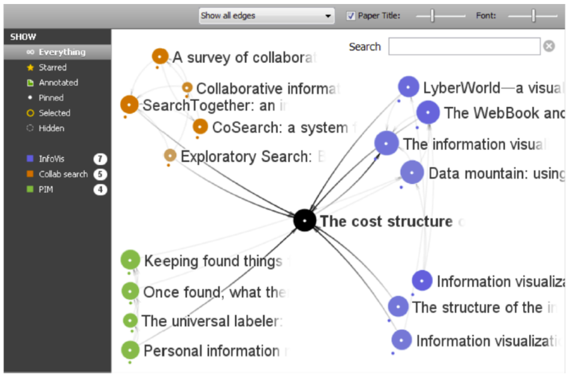


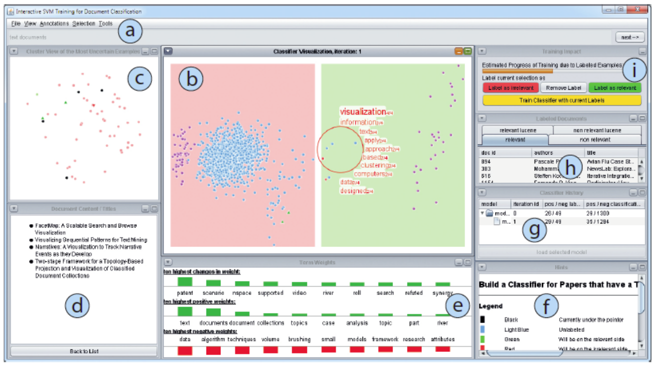
Configuration
Target
Name: Test
Value:
Model
Algorithm: CRF
Settings: Binary Vectors (Bin); Overlapping Features (Ovlp); Case Insensitive (ICase)
Features
Training Unit: JudgingAnnotation
Testing Unit: NP
Output Unit: NP
Features: WordToken, UmlsConcept#cui, WordToken#canonicalForm, NegatedCui#negatedCui, DateAnnotation, NamedEntity
Validation
Mode: K-Fold
Settings: 10 folds

Results
Rows: Features
Columns: Testing Unit
F-Measure | Recall | Precision | Combined

<table>
<tr><th></th><th>NP</th><th>Annota...</th></tr>
<tr><td>UmlsConcept#cui, AnnotationContext</td><td>[illegible]</td><td>☐ Run</td></tr>
<tr><td>NumToken#lemmaEntries</td><td>0</td><td>0</td></tr>
<tr><td>INTJ</td><td>0</td><td>0</td></tr>
<tr><td>WordToken, RomanNumeralAnnotation, UmlsConcept#cui, WordToken#canonicalF...</td><td>☑ Run</td><td>0.8362...</td></tr>
<tr><td>NamedEntity#typeID</td><td>0.2029...</td><td>0.1988...</td></tr>
<tr><td>NamedEntity#status</td><td>0</td><td>0</td></tr>
<tr><td>UmlsConcept#cui, AnnotationContext, UmlsConcept#code</td><td>0.7038...</td><td>☑ Run</td></tr>
<tr><td>NumToken#partOfSpeech</td><td>0</td><td>0</td></tr>
<tr><td>WordToken, UmlsConcept#code</td><td>0.8246...</td><td>0.8226...</td></tr>
<tr><td>MeasurementAnnotation, NamedEntity#certainty, NENegation, AnnotationContext,...</td><td>0.7752...</td><td>0.8155...</td></tr>
<tr><td>UmlsConcept#cui, WordToken#canonicalForm</td><td>0.8222...</td><td>0.8221...</td></tr>
<tr><td>PP</td><td>0</td><td>0</td></tr>
<tr><td>NegatedCui#negatedCui</td><td>0.0195...</td><td>0.0233...</td></tr>
<tr><td>MeasurementAnnotation</td><td>0</td><td>0</td></tr>
<tr><td>WordToken, AnnotationContext, UmlsConcept#code</td><td>0.8267...</td><td>☐ Run</td></tr>
<tr><td>NP, UmlsConcept#cui, UmlsConcept#code, WordToken#canonicalForm</td><td>0.8401...</td><td>☐ Run</td></tr>
<tr><td>AnnotationContext</td><td>0.1996...</td><td>0.2256...</td></tr>
<tr><td>CONJP</td><td>0</td><td>0</td></tr>
<tr><td>WordToken, RomanNumeralAnnotation, UmlsConcept#cui, UmlsConcept#code</td><td>☐ Run</td><td>0.8247...</td></tr>
<tr><td>LymphLevelExpression#numTotal</td><td>0</td><td>0</td></tr>
<tr><td>NEContext#FocusText</td><td>0</td><td>0.0061...</td></tr>
<tr><td>PunctuationToken#normalizedForm</td><td>0</td><td>0</td></tr>
<tr><td>UmlsConcept#oid</td><td>0</td><td>0</td></tr>
<tr><td>RomanNumeralAnnotation, UmlsConcept#cui, UmlsConcept#code</td><td>☐ Run</td><td>0.6909...</td></tr>
<tr><td>LymphLevelExpression#numPositive</td><td>0</td><td>0</td></tr>
<tr><td>UmlsConcept#cui, AnnotationContext, WordToken#canonicalForm</td><td>0.8294...</td><td>☐ Run</td></tr>
<tr><td>SizeExpression</td><td>0</td><td>0</td></tr>
<tr><td>PersonTitleAnnotation</td><td>0</td><td>0</td></tr>
<tr><td>UmlsConcept#codingScheme</td><td>0</td><td>0</td></tr>
<tr><td>WordToken, UmlsConcept#cui, UmlsConcept#code</td><td>0.8241...</td><td>0.8243...</td></tr>
<tr><td>SBAR</td><td>0</td><td>0</td></tr>
<tr><td>LST</td><td>0</td><td>0.0010...</td></tr>
<tr><td>FractionAnnotation</td><td>0</td><td>0</td></tr>
<tr><td>NamedEntity#uid</td><td>0</td><td>0</td></tr>
<tr><td>SymbolToken#normalizedForm</td><td>0</td><td>0</td></tr>
<tr><td>AnnotationContext, UmlsConcept#code</td><td>0.6980...</td><td>☐ Run</td></tr>
</table>
</td>
</tr>
<tr>
<td><em>Information Extraction & Pattern Matching</em>

Canary [12]</td>
<td>Graphical interface for information extraction using rule-based NLP with user-defined grammars and lexicons. Canary provides GUIs for different steps of the pipeline, such as text normalization, phrase structure rules, etc. However, it does require the help of an expert who can understand these steps for developing a model.</td>
</tr>
</table>

2. *End-user focused tools* for users with little to no machine learning experience	
Task / Publication	**Description**
Clustering Apolo [26]	**Review:** Apolo shows citations of an article clustered into different groups. Each citation is assigned to a color-coded cluster with color saturation representing belongingness. Some articles can be marked as exemplars for a specific cluster. Users can manually inspect articles. **Feedback:** Allows clusters to be manually added and removed. Users may also specify exemplar articles to form the basis for the clusters for improving the models. The tool offers an example of how interactive NLP system may be used for building document clusters. It also presents the idea of using exemplars for receiving user feedback along with other innovations in the visual exploration of the clusters predicted by the machine learning algorithm.  Show all edges ☑ Paper Title: Font: Search SHOW Everything Starred Annotated Pinned Selected Hidden InfoVis 7 Collab search 5 PIM 4 A survey of collabora Collaborative informa SearchTogether: an i CoSearch: a system Exploratory Search: E LyberWorld—a visua The WebBook and The information visua Data mountain: using The cost structure Keeping found things Once found, what the The universal labeler Personal information Information visualiza The structure of the i Information visualizati

Classification Visual Classifier Training [27]	**Review:** Presents an interactive visualization that projects the document into a 2D plot along with the decision boundary. Users can explore the word cloud of individual documents. **Feedback:** Uses both active learning based as well as user-steered workflows for receiving new document labels. The interface presents a visualization of most uncertain documents. **Retrain:** Visualizes estimated feedback impact, training progress, model weights, etc. It also shows a bar-chart visualization of features weights that undergo most change upon model revisions. Although the tool may be used by those without machine learning experience, it is still inclined towards data-scientists. For applications in the clinical domain, we will also need to consider the design requirements of clinical tasks in hand and the objectives for which models are being built for. 

Interactive SVM Training for Document Classification
File View Annotations Selection Tools
a
next -->
Cluster View of the Most Uncertain Examples
c
Classifier Visualization, iteration: 1
b
visualization
information
text
apply
approach
based
clustering
computers
data
designed
Training Impact
Estimated Progress of Training due to Labeled Examples
i
Label current selection as
Label as irrelevant
Remove Label
Label as relevant
Train Classifier with current Labels
Labeled Documents
relevant lucene
non relevant lucene
relevant
non relevant
doc id
authors
title
894
383
516
Pascale F
Mohamm
Steffen Ko
Avian Flu Case St..
NewsLab: Explora..
Iterative Integratio..
h
Classifier History
model
iteration id
pos / neg lab...
pos / neg classificati...
mod.. 0
26 / 49
29 / 1300
m... 1
29 / 49
36 / 1284
g
load selected model
Document Content / Titles
FacetMap: A Scalable Search and Browse Visualization
Visualizing Sequential Patterns for Text Mining
Narratives: A Visualization to Track Narrative Events as they Develop
Two-stage Framework for a Topology-Based Projection and Visualization of Classified Document Collections
d
Back to List
Term Weights
ten highest changes in weight:
patent
scenario
nspace
supported
video
river
roll
search
refuted
synergy
ten highest positive weights:
text
documents
document
collections
topics
case
analysis
topic
part
river
ten highest negative weights:
data
algorithm
techniques
volume
brushing
small
models
framework
research
attributes
e
Hints
Build a Classifier for Papers that have a T
Legend
Black
Currently under the pointer
Light Blue
Unlabeled
Green
Will be on the relevant side
Red
Will be on the irrelevant side
f

<table>
<tr>
<td><em>Classification</em>

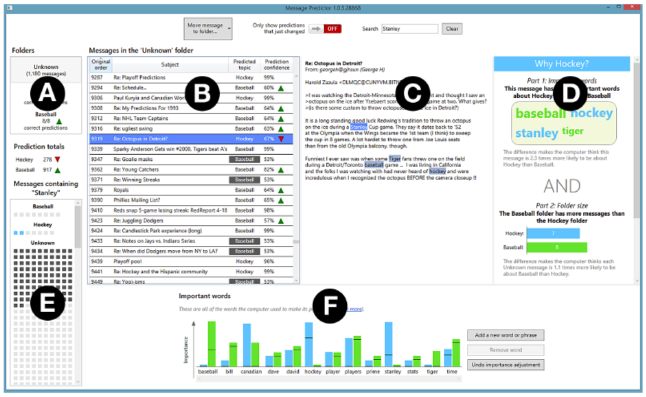
EluciDebug [37]</td>
<td>

Review: Similar to NLPReViz [48] and allows users to review binary classification models. It highlights the top features in the document and shows visualization of different feature weights across multiple views.

Feedback: User-feedback includes providing new labels for whole documents, as well as adding or removing features from the learning model. Alternatively, the users are also able to alter the feature weights in the model. The focus of the tool is to enable end-users to debug the models but lacks capabilities of making the feedback in more usable and efficient manners.

Retrain: Uses explanatory debugging to encourage users to correct the models by providing feedback in an iterative manner. All these changes are tracked and can be easily reversed. It also shows how the prediction confidence changes for individual documents over successive model revisions.

Similar to [27], EluciDebug is primarily focused on the task of model building and exploration. There is scope to design requirements defined by the individual clinical tasks and making the feedback mechanisms more efficient.



Message Predictor 1.0.5.28868
Move message to folder...
Only show predictions that just changed
OFF
Search
Stanley
Clear
Folders
Unknown (1,180 messages)
Baseball 8/8 correct predictions
Prediction totals
Hockey 278
Baseball 917
Messages containing "Stanley"
Baseball
Hockey
Unknown
Messages in the 'Unknown' folder
A
B
C
D
E
F

<table>
<tr><th>Original order</th><th>Subject</th><th>Predicted topic</th><th>Prediction confidence</th></tr>
<tr><td>9287</td><td>Re: Playoff Predictions</td><td>Hockey</td><td>99%</td></tr>
<tr><td>9294</td><td>Re: Schedule...</td><td>Baseball</td><td>60% ▲</td></tr>
<tr><td>9306</td><td>Paul Kuryia and Canadian Wor</td><td>Hockey</td><td>99%</td></tr>
<tr><td>9308</td><td>Re: My Predictions For 1993</td><td>Baseball</td><td>64% ▲</td></tr>
<tr><td>9312</td><td>Re: NHL Team Captains</td><td>Baseball</td><td>64% ▲</td></tr>
<tr><td>9316</td><td>Re: ugliest swing</td><td>Baseball</td><td>63% ▲</td></tr>
<tr><td>9319</td><td>Re: Octopus in Detroit?</td><td>Hockey</td><td>67% ▼</td></tr>
<tr><td>9339</td><td>Sparky Anderson Gets win #2000, Tigers beat A's</td><td>Baseball</td><td>99%</td></tr>
<tr><td>9347</td><td>Re: Goalie masks</td><td>Baseball</td><td>53%</td></tr>
<tr><td>9362</td><td>Re: Young Catchers</td><td>Baseball</td><td>82% ▲</td></tr>
<tr><td>9371</td><td>Re: Winning Streaks</td><td>Baseball</td><td>53%</td></tr>
<tr><td>9379</td><td>Royals</td><td>Baseball</td><td>64% ▲</td></tr>
<tr><td>9390</td><td>Phillies Mailing List?</td><td>Baseball</td><td>65% ▲</td></tr>
<tr><td>9410</td><td>Reds snap 5-game losing streak: RedReport 4-18</td><td>Baseball</td><td>98%</td></tr>
<tr><td>9423</td><td>Re: Juggling Dodgers</td><td>Baseball</td><td>57% ▲</td></tr>
<tr><td>9424</td><td>Re: Candlestick Park experience (long)</td><td>Baseball</td><td>99%</td></tr>
<tr><td>9433</td><td>Re: Notes on Jays vs. Indians Series</td><td>Baseball</td><td>53%</td></tr>
<tr><td>9434</td><td>Re: When did Dodgers move from NY to LA?</td><td>Baseball</td><td>53%</td></tr>
<tr><td>9439</td><td>Playoff pool</td><td>Hockey</td><td>96%</td></tr>
<tr><td>9441</td><td>Re: Hockey and the Hispanic community</td><td>Hockey</td><td>99%</td></tr>
<tr><td>9449</td><td>Re: Yogi-isms</td><td>Baseball</td><td>53%</td></tr>
</table>

Re: Octopus in Detroit?
From: *georgeh@gjhsun (George H)*

Harold Zazula <DLMQC@CUNYVM.BITN

>I was watching the Detroit-Minnesota nt and thought I saw an
>octopus on the ice after Ysebaert scor game at two. What gives?
>(is there some custom to throw octopus ice in Detroit?)

It is a long standing good luck Redwing's tradition to throw an octopus on the ice during a Stanley Cup game. They say it dates back to '52 at the Olympia when the Wings became the 1st team (I think) to sweep the cup in 8 games. A lot harder to throw one from Joe Louis seats than from the old Olympia balcony, though.

Funniest I ever saw was when some Tiger fans threw one on the field during a Detroit/Toronto baseball game ... I was living in California and the folks I was watching with had never heard of hockey and were incredulous when I recognized the octopus BEFORE the camera closeup !!

Why Hockey?

Part 1: Im words
This message has tant words about Hockey Baseball

baseball hockey stanley tiger

The difference makes the computer think this message is 2.3 times more likely to be about Hockey than Baseball.

AND

Part 2: Folder size
The Baseball folder has more messages than the Hockey folder

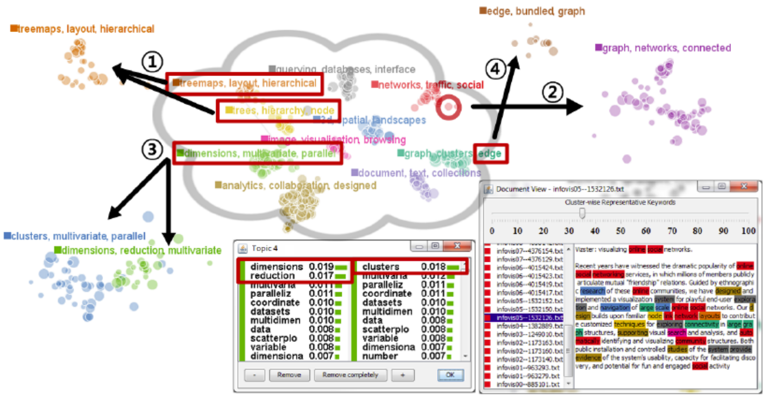
Hockey: 7
Baseball: 8

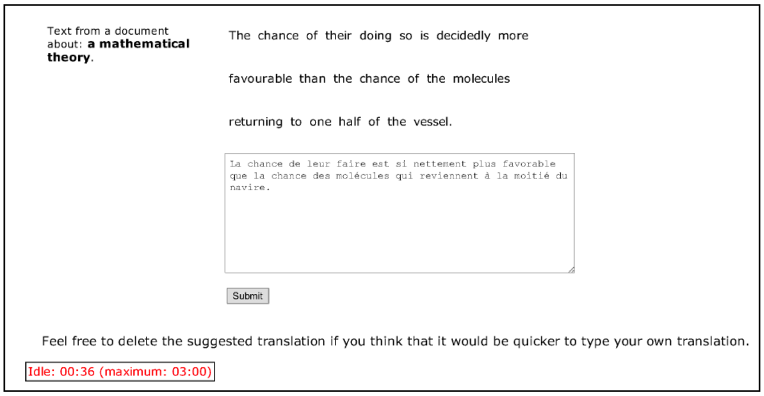
The difference makes the computer thinks each Unknown message is 1.1 times more likely to be about Baseball than Hockey.

Important words

These are all of the words the computer used to make its p more!

</td>
</tr>
</table>

Topic Modeling UTOPIAN [59]	**Review:** The system provides a simple visualization of topic models using word clouds and shows keyword highlights in full-text. **Feedback:** Users can provide feedback by merging topics, induce new topics by documents and keywords, and also split existing topics. They may adjust the weights for topic keywords as well. Feedback is supported by a semi-supervised algorithm proposed by the authors in this work. **Retrain:** The tool supports retraining, but does not provide easy interfaces for users to understand changes between model revisions, switch between different model versions, etc.  edge, bundled graph treemaps, layout, hierarchical graph, networks, connected ① ④ querying, databases, interface treemaps, layout, hierarchical networks, traffic, social ② trees, hierarchy, node 3d, spatial, landscapes image, visualisation, browsing ③ dimensions, multivariate, parallel graph, clusters, edge document, text, collections analytics, collaboration, designed clusters, multivariate, parallel dimensions, reduction, multivariate Topic 4 - Remove Remove completely + OK Document View - infovis05--1532126.txt Cluster-wise Representative Keywords 0 10 20 30 40 50 60 70 80 90 100 dimensions 0.019 reduction 0.017 multivaria 0.011 paralleliz 0.011 coordinate 0.010 datasets 0.010 multidimen 0.010 data 0.008 scatterplo 0.008 variable 0.008 dimensiona 0.007 clusters 0.018 multivaria 0.012 paralleliz 0.011 coordinate 0.011 datasets 0.010 multidimen 0.010 data 0.008 scatterplo 0.008 variable 0.008 dimensiona 0.007 number 0.007 infovis07--4376154.txt infovis07--4376129.txt infovis06--4015424.txt infovis06--4015423.txt infovis06--4015419.txt infovis06--4015417.txt infovis05--1532152.txt infovis05--1532150.txt infovis05--1532126.txt infovis04--1382889.txt infovis03--1249010.txt infovis02--1173163.txt infovis02--1173160.txt infovis02--1173140.txt infovis01--963293.txt infovis01--963279.txt infovis00--885101.txt Vizster: visualizing online social networks. Recent years have witnessed the dramatic popularity of online social networking services, in which millions of members publicly articulate mutual "friendship" relations. Guided by ethnographic research of these online communities, we have designed and implemented a visualization system for playful end-user exploration and navigation of arge scale online social networks. Our design builds upon familiar node ink network layouts to contribute customized techniques for exploring connectivity in arge graph structures, supporting visual search and analysis, and automatically identifying and visualizing community structures. Both public installation and controlled studies of the system provide evidence of the system's usability, capacity for facilitating discovery, and potential for fun and engaged social activity

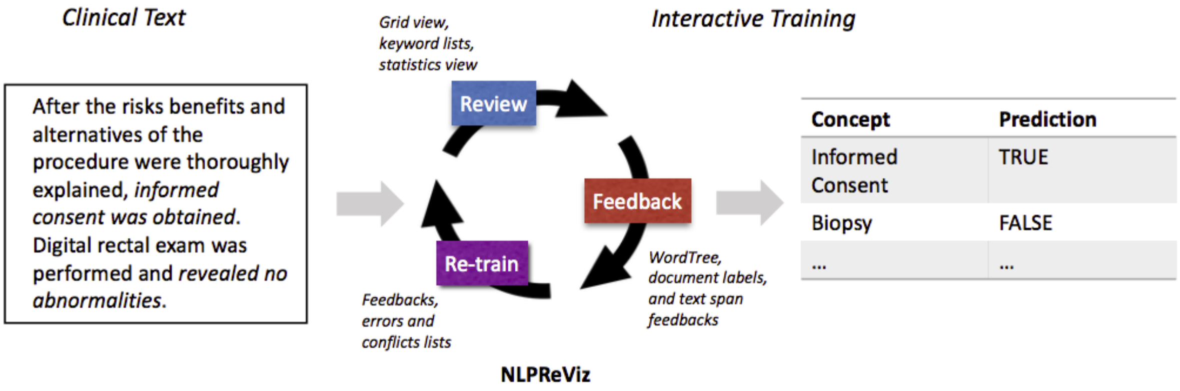
Topic Modeling Dissertation Browser [60]	**Review:** Similar to UTOPIAN [59], but focuses on making topic models 'interpretable' and 'trustworthy' using visualization support. It implements richer interactions as compared to UTOPIAN, for example, it provides support to compare different models by visualizing the similarity between topics, allows multiple zoom-levels to inspect the full-text data while inspecting the models, etc. **Feedback:** Recommends some ideas for feedback and retraining for future work, but does not implement them.
Translation Human Post-Editing [61]	**Review:** Allows manual correction of machine translated text by humans. The results show that such an approach leads to a reduced time as well as an improvement in the quality of translations. Provides visualization support to draw users' attention to individual words and phrases that need more attention. **Feedback:** Models are not revised based on human feedback. Only the corrections against individual translations are retained.  Text from a document about: **a mathematical theory**. The chance of their doing so is decidedly more favourable than the chance of the molecules returning to one half of the vessel. La chance de leur faire est si nettement plus favorable que la chance des molécules qui reviennent à la moitié du navire. Submit Feel free to delete the suggested translation if you think that it would be quicker to type your own translation. Idle: 00:36 (maximum: 03:00)

Many of these tools described in Table 3 only partially address the three steps of the interactive learning and lack end-to-end implementation of the cycle. For example, tools such as LightSIDE [57] and ARC [53] provide GUI supports for the users to tweak the parameters to the learning model. Other tools designed for non-machine learning expert users, focus only on specific interface components. For example, RapTAT [51, 52] provides support for building annotated training sets but do not provide an interactive cycle. Human-post editing for machine translation [61] and Dissertation Browser [60] help only with the review and feedback steps. They suggest but do not implement or evaluate ideas for building model revisions and retraining. Even the tools that cover the entire learning cycle, such as Visual Classifier Training [27] and EluciDebug [37], are designed for very specific NLP tasks and use-cases like document classification.

4.0 *NLPReViz:* INTERACTIVE NLP FOR RETROSPECTIVE REVIEW

I first demonstrate the interactive NLP approach on clinical text for retrospective research. I picked an example problem of classifying colonoscopy procedure notes to generate measures for a quality improvement program [62, 63]. In order to analyze these records with the existing tools, researchers must either go through expensive NLP model building process involving data-scientists or manually read through the records to extract the information of interest. To address this problem, we built NLPReViz [48] – an interactive web-based tool designed for allowing clinicians and clinical researchers to interactively build NLP models (Figure 5). In this chapter, I describe the design, implementation, and evaluation of NLPReViz. This also serves as a reference framework for proceeding with the work in the next chapter (Chapter 5).

Our user interface design complements the rationale based learning system that we adopted to incorporate user feedback [64]. This approach allows domain experts without machine learning experience to build models and give feedback to improve them iteratively. To support this model, NLPReViz incorporates three different kinds of feedback mechanisms: a) assigning labels for the whole document, b) selecting text spans indicative of a specific value, or c) selecting phrases found across multiple documents using the WordTree [65] as shown in Figure 7(a). The WordTree is useful for providing feedback on several documents together in an efficient manner as described detail in the following sections. We also conducted user studies supporting the viability of our approach by demonstrating notable improvements in performance metrics in a short time span, with minimal initial training.



After the risks benefits and alternatives of the procedure were thoroughly explained, *informed consent was obtained*. Digital rectal exam was performed and *revealed no abnormalities.*

Concept	Prediction
Informed Consent	TRUE
Biopsy	FALSE
...	...

Figure 5: **An interactive machine learning cycle begins with the review step, with the output from the learning model displayed to the user. User feedback is used by the system to improve upon the machine learning models by providing labels for documents that were previously not part of the training set, or by correcting any misclassified documents. After re-training, a new model is created, and the tool highlights show prediction changes along with providing guidance for resolving potentially contradictory feedback items.**

4.1 LEARNING MODEL

We use bag of words and Support Vector Machine classifiers (SVMs) with linear kernels to predict binary classifications for concept variables extracted from documents. Our model for incorporating user feedback adapts a framework proposed by Zaidan et al. [64], in which domain experts supply not only the correct label but also a span of text that serves as a rationale for their labeling decision. Rationales are turned into pseudo-examples providing additional training data [64, 66]. Rationales have been shown to be effective for predicting sentiments of movie reviews [66]. We adapted this approach for use on clinical text by constructing one merged pseudo example per document from the annotations received. Rationales are constructed from user interactions with the tool and are used to retrain the SVM models.

4.2 INTERFACE DESIGN

The design of our tool was informed by prior work on interactive learning systems [34, 22, 67]. These design requirements for it can be divided into three according to the interactive learning cycle (Figure 3):

(i) **Review** displays support the interpretation of NLP results both within and across documents.

R1: Document displays highlight NLP results and, where possible, show evidence for the results extracted from the text.

R2: Overview displays support comparison between documents and identification of frequent words or phrases associated with NLP results.

(ii) **Feedback** mechanisms provide usable and efficient means of updating NLP models.

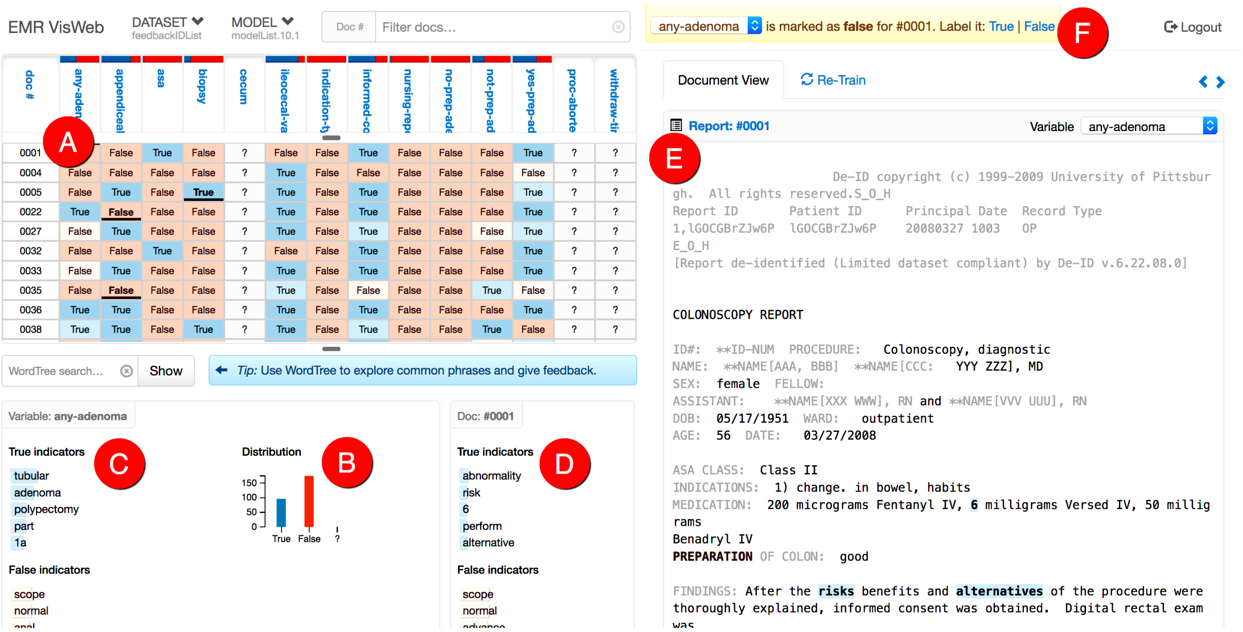
R3: Interaction tools support the selection of text as evidence for selected interpretations.

R4: Conflicting or inconsistent feedback should be identified and presented to the user for appropriate resolution.

(iii) **Re-train** Results of model revisions should be apparent to users.

R5: Displays should help the users understand changes in predictions and other model revisions.

Figures 6 and 7 show the different components of NLPReViz's user interface. A video demo can be found at `vimeo.com/trivedigaurav/emr-demo`. We built a prototype, evaluated it with a think-aloud study, and revised it based on the participants' feedback [62]. Our tool is available for download along with source code and documentation at `NLPReViz.github.io`.



doc #	any-aden	appendiceal	asa	biopsy	cecum	ileocecal-va	indication-ty	informed-co	nursing-repo	no-prep-ade	not-prep-ad	yes-prep-ad	proc-aborte	withdraw-tir
0001	[illegible]	False	True	False	?	False	False	True	False	False	False	True	?	?
0004	False	False	False	False	?	True	False	False	False	False	False	False	?	?
0005	False	True	False	**True**	?	True	False	True	False	False	False	True	?	?
0022	True	**False**	False	False	?	True	False	True	False	False	False	True	?	?
0027	False	True	False	False	?	True	False	True	False	False	False	True	?	?
0032	False	False	True	False	?	False	False	True	False	False	False	True	?	?
0033	False	True	False	False	?	True	False	True	False	False	False	True	?	?
0035	False	**False**	False	False	?	True	False	False	False	False	True	False	?	?
0036	True	True	False	False	?	True	False	True	False	False	False	True	?	?
0038	True	True	False	True	?	True	False	True	False	False	True	False	?	?

```
                         De-ID copyright (c) 1999-2009 University of Pittsbur
gh.  All rights reserved.S_O_H
Report ID        Patient ID       Principal Date  Record Type
1,lGOCGBrZJw6P  lGOCGBrZJw6P     20080327 1003   OP
E_O_H
[Report de-identified (Limited dataset compliant) by De-ID v.6.22.08.0]


COLONOSCOPY REPORT

ID#:  **ID-NUM  PROCEDURE:   Colonoscopy, diagnostic
NAME:  **NAME[AAA, BBB]  **NAME[CCC:   YYY ZZZ], MD
SEX:  female  FELLOW:
ASSISTANT:   **NAME[XXX WWW], RN and **NAME[VVV UUU], RN
DOB:  05/17/1951  WARD:   outpatient
AGE:  56  DATE:   03/27/2008

ASA CLASS:  Class II
INDICATIONS:  1) change. in bowel, habits
MEDICATION:  200 micrograms Fentanyl IV, 6 milligrams Versed IV, 50 millig
rams
Benadryl IV
PREPARATION OF COLON:  good

FINDINGS: After the risks benefits and alternatives of the procedure were
thoroughly explained, informed consent was obtained.  Digital rectal exam
was
```

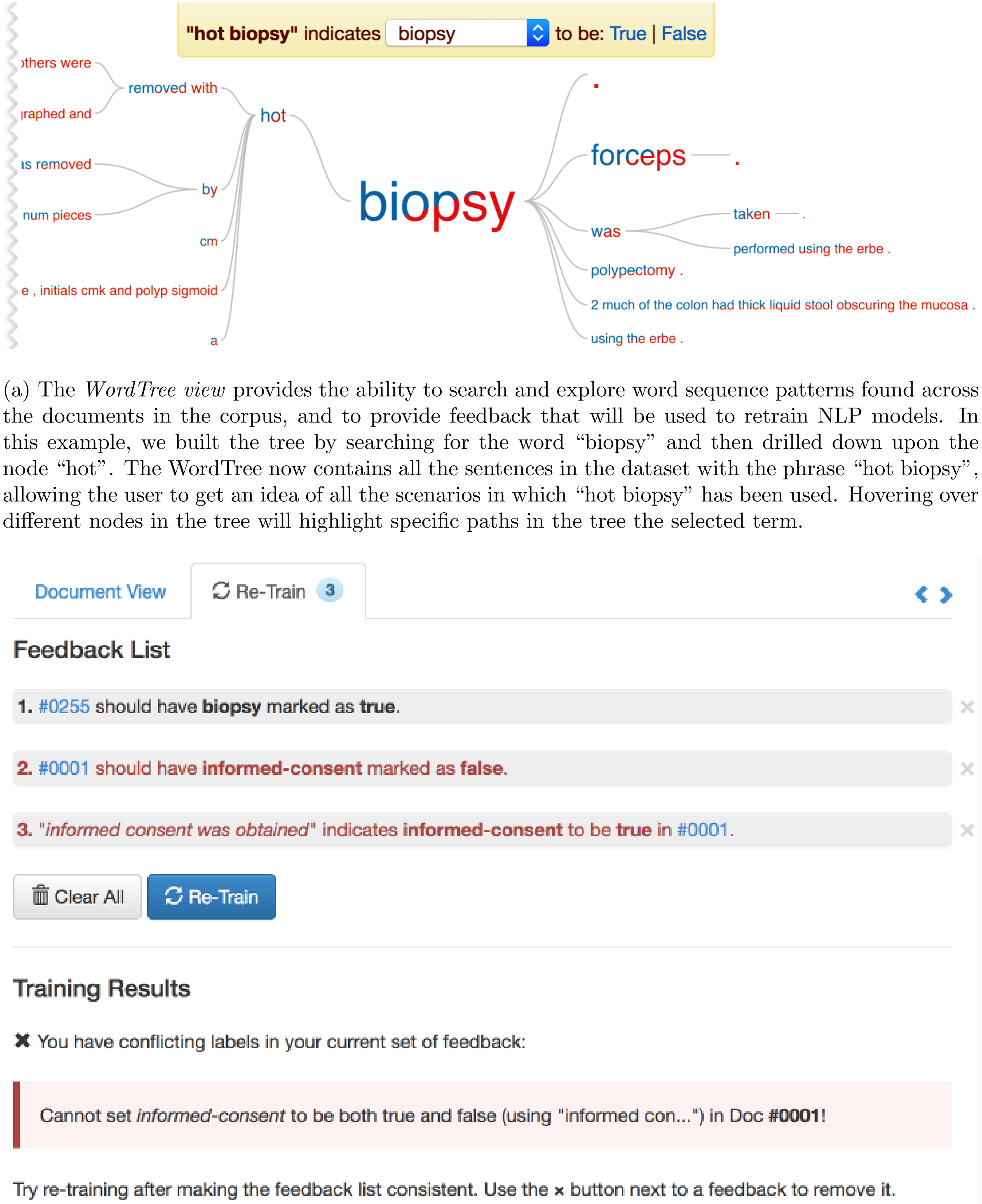
Figure 6: **(a) The *Grid view* shows the extracted variables in columns and individual documents in rows, providing an overview of NLP results. Below the grid, we have statistics about the active variable with (b) the distribution of the classifications for the selected variable and (c) the list of top indicators for that variable aggregated across all the documents in the dataset. (d) Indicators from the active report are shown on the right. (e) The *document view* shows the full-text of the patient reports with the indicator terms highlighted. (f) Feedback can be sent using the yellow control bar on the top, or by using a right-click context menu.**

4.3 EVALUATION

Our evaluation addressed two key questions [48]: 1) can clinicians successfully use NLPReViz to provide feedback for improving NLP models, and 2) can this feedback be effective with a small set of initial training data?

4.3.1 Dataset

We used a reduced dataset of colonoscopy reports prepared by Harkema et al. [68] along with their gold standard label set. Participants worked with two variables: ‘biopsy’ and ‘appendiceal-orifice’. A document was marked true for the biopsy variable if the report


"hot biopsy" indicates biopsy to be: True | False
others were
removed with
graphed and
hot
as removed
by
num pieces
cm
e , initials cmk and polyp sigmoid
a
biopsy
.
forceps .
was
taken .
performed using the erbe .
polypectomy .
2 much of the colon had thick liquid stool obscuring the mucosa .
using the erbe .

(a) The *WordTree view* provides the ability to search and explore word sequence patterns found across the documents in the corpus, and to provide feedback that will be used to retrain NLP models. In this example, we built the tree by searching for the word "biopsy" and then drilled down upon the node "hot". The WordTree now contains all the sentences in the dataset with the phrase "hot biopsy", allowing the user to get an idea of all the scenarios in which "hot biopsy" has been used. Hovering over different nodes in the tree will highlight specific paths in the tree the selected term.

Feedback List

1. #0255 should have **biopsy** marked as **true**. ×

2. #0001 should have **informed-consent** marked as **false**. ×

3. "*informed consent was obtained*" indicates **informed-consent** to be **true** in #0001. ×

Training Results

✖ You have conflicting labels in your current set of feedback:

Cannot set *informed-consent* to be both true and false (using "informed con...") in Doc **#0001**!

Try re-training after making the feedback list consistent. Use the × button next to a feedback to remove it.

(b) The *Re-Train view* lists user-provided feedback, including any potential inconsistencies, and specifies changes in variable assignments due to retraining. In the example above, the user has selected a text span documenting "informed-consent" in a report. However, they also labeled the report incorrectly, possibly in error. NLPReViz points this out as conflicting feedback.

Figure 7: **Screenshots of the *WordTree* and *Re-Train* views.**

indicated that a sample of tissue was tested through a biopsy procedure. The appendiceal-orifice variable indicates whether that region of the colon was reached and was explicitly noted during the colonoscopy. Our dataset consisted of 453 documents, split into two parts: two-thirds for a *development set* for conducting the user study, and one-third held out as a *test set* for evaluating the system performance.

4.3.2 Participants

We identified a convenience sample of participants with MD degrees and knowledge of colonoscopy procedures. Participants were given a $50 gift card for 90 minutes of participation via web conferencing. One participant (p9) experienced technical difficulties resulting in shorter study time. To address the question of sensitivity to the size of the initial training set, we used two splits to build initial training models. The first group of four participants (p1-p4) started with models built on 10 annotated documents. Initial models for the second group (p5-p8 and p9) were based on 30 annotated documents. The same 173 documents were used in the test set for both groups.

4.3.3 Protocol

Each session began with a participant background questionnaire, followed by a 15-minute walk-through of the interface and an introduction to the annotation guidelines used for preparing our gold standard labels.] Participants were given up to one hour to annotate and build models, roughly divided between the two variables. We reminded them to retrain at regular intervals, particularly if they provided more than 10 consecutive feedback items without retraining. After finishing both variables, participants completed the System Usability Scale [69] and discussed reactions to the tool. We evaluated the performance of the models on the test set using the harmonic mean of recall and precision – F1 score at each retraining step. We calculated Cohen's κ statistic [70] to measure the agreement of the complete set of each participants feedback items with the gold standard labels. To compare user feedback to a possibly optimal set of labels, we simulated feedback actions using gold standard labels. 10 random feedback items (without rationales) were added at each step, ranging from 10-280 items. This was repeated 50 times to compute an average.

We used a reduced dataset of colonoscopy reports prepared by Harkema et al. [68] along with their gold standard label set. Our dataset consisted of 453 documents, split into two parts: two-thirds for a development set for conducting the user study, and one-third held out as a test set for evaluating the system performance. Participants worked with two variables: 'biopsy' and 'appendiceal-orifice'. Each session began with a participant background questionnaire, followed by a 15-minute walkthrough of the interface and an introduction to the annotation guidelines used for preparing our gold standard labels. Participants were given up to one hour to annotate and build models, roughly divided between the two variables. After finishing both variables, participants completed the System Usability Scale [69] and discussed reactions to the tool.

We evaluated the performance of the models on the test set using the harmonic mean of recall and precision – F1 score at each retraining step. We calculated Cohens κ statistic to measure the agreement of the complete set of each participant's feedback items with the gold standard labels. To compare user feedback to a possibly optimal set of labels, we simulated feedback actions using gold standard labels. Random feedback items (without rationales) were added at each step, ranging from 10-280 items. This was repeated 50 times to compute an average.

4.4 RESULTS

Nine physicians participated in our study. The average SUS score was 70.56 out of 100. A SUS score of 68 is considered as average usability [69]. The changes in F1 scores on the test set (relative to gold-standard labels) for the nine participants are shown in Figure 8, along with their Cohen's κ scores indicating agreement of feedback with the gold-standard labels. Scores are plotted against the cumulative number of records affected by user feedback actions after each retraining step. Performance improved in 17 of 18 tasks, with improvements as high as 29.90%.

We found improvements in F1 scores across all users for the appendiceal-orifice, though results were more mixed for biopsy. Examination of less successful efforts indicated that some participants found the biopsy annotation guidelines to be challenging. These difficulties were associated with the lower kappa scores (eg. p8, biopsy) between the user provided labels and gold-standard labels. Favorable performance of models based on participant feedback, relative to results of simulations using gold standard labels (*right-half* of Figure 8) suggests that NLPReViz can be used to elicit feedback suitable for improving NLP models. The differences in participants' approach for annotation and retraining are summarized in Table 4.

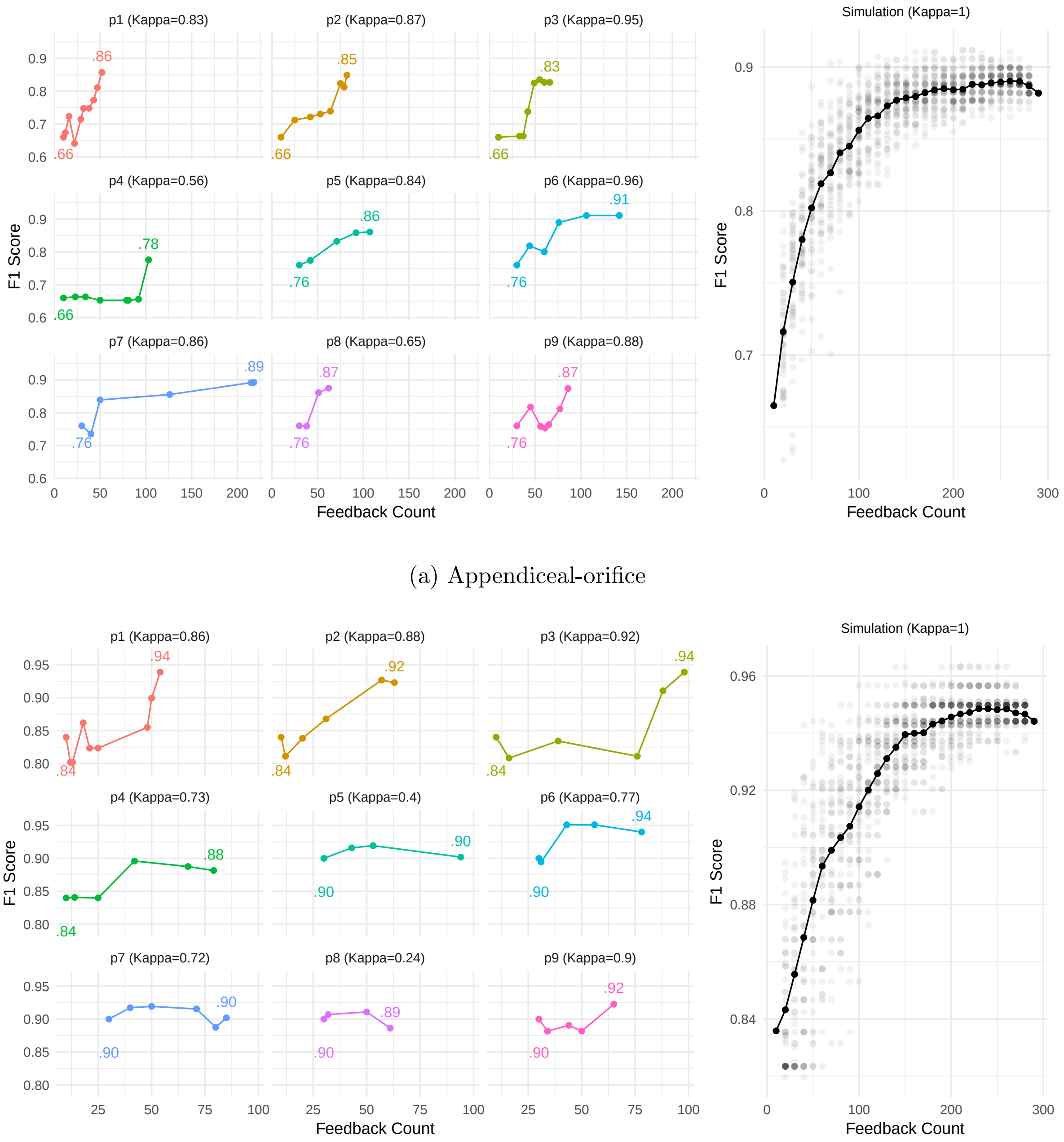
Open-ended subjective feedback was generally positive toward the design of the tool. Participants commented on the overall design: *"The system's functions were very well integrated – I think it was very nice"*, *"It was very well thought out: the WordTree was beautiful – the reds and the blue"* and *"I'd be happy to use the tool more often."* Others commented about the learnability of the tool: *"I thought it was very easy to use and straightforward"*, *"The process was very easy with a little bit of guidance"*, and *"May need some initial training - may be complex for somebody who hasn't done [annotations] before."*. Comments regarding desired additional functionality stressed the need for clearer indications of which documents had been labeled, navigational shortcuts, classification of text spans as irrelevant to a given classification, improvements to the retraining process, and other enhancements.

4.5 DISCUSSION

We developed a prototype tool for helping clinicians build models for binary concepts on their own. NLPReViz combines interactive displays of NLP results with tools for finding patterns of interest, reviewing text, and revising NLP models. Eliciting clinician feedback for review and revision of NLP models requires a combination of views for displaying documents and NLP results in context with means of providing feedback required to revise the models. Our user study demonstrated successful use of the tool on small data set, raising the possibility of constructing NLP models with minimal training. This initial success with small training

Table 4: ***Activity Pattern*: This table summarizes the activity patterns of the individual participants. The *p1-p4* started with an initial model trained on 10 documents, while *p5-p9* started with 30 documents.**

	Docs Opened	**Total Feedback**	**Unique Feedback**	**Unseen Feedback**	**Model Count**	**Error Count**	**Type of Feedback**		
							WordTree	**Span**	**Label**
p1	129	86	66	14	17	2	1 (19)	16	51
p2	117	125	112	1	12	6	4 (40)	24	61
p3	71	144	88	5	12	3	2 (16)	28	98
p4	94	162	93	4	12	1	6 (43)	26	106
p5	104	141	133	14	7	1	1 (18)	6	117
p6	170	301	230	29	6	2	3 (44)	58	181
p7	50	243	202	141	6	2	8 (190)	21	32
p8	54	63	55	0	6	1	0 (0)	23	40
p9	68	91	81	0	10	2	0 (0)	34	57
	Combined count of documents reviewed for both variables	Number of feedback items provided	Unique documents labels inferred from feedback	Documents labeled without viewing them first (when using the WordTree)	Number of training iterations	Conflicts and overrides in provided feedback	Feedback items provided using the different feedback input mechanisms: the WordTree view (along with the documents affected), highlighting spans or assigning a label to the document		



(a) Appendiceal-orifice

(b) Biopsy

Figure 8: **Plots showing the variations in F1-Scores for the two variables as the participants provided feedback. These results are shown for the test dataset only. Participants p1-p4 start with an initial training set of 10 documents, while p5 onward used a model trained on 30 training documents. Differences in the spacing of the points in each graph reflect differences in feedback rates across participants. Kappa scores next to the participant ids indicate how well their feedback compared to the gold standard labels.**

sets suggests the possibility of using our approach to develop annotation guidelines *de novo*, perhaps with pre-annotation techniques similar to those used in RapTAT [51, 52].

Our interface and interaction design for reviewing documents and models were well received by the participants. They found it useful to review keyword-based features highlighted in individual documents which could be quickly identified using our color scheme. However, views displaying overall class distribution and dataset-level feature list remained little-used parts of the interface. We think that this may be because they were less familiar with the dataset being reviewed and didn't have any prior baselines to compare them to. There may also be further scope to better refine these displays. They were able to make use of all three types of feedback mechanisms for revising models. Although they showed variations in individual preference between them (Table 4), they also discovered an unexpected use case for this view. In addition to giving feedback, the WordTree allowed users to get a sense of the quality of their models. This was a consequence of the gradient colors, which showed how the presence of individual keywords affects the classification of documents. User feedback also suggested several possible improvements to our design. One particularly interesting idea involved the need to indicate that a phrase was irrelevant to a classification of a document. This would allow the user to remove non-informative but possibly misleading features for model-building. This might have proven useful for the biopsy variable, where some participants may have been confused by the presence of the term "hot biopsy", which indicated a tool used to remove polyps and not a biopsy procedure. The retrain views allowed the users to see changes in predictions after model revisions. However, it lacked sufficient mechanisms for estimating model performance over time. More specifically, we did little to address the problem of letting the users know when the model is good enough and the provided number of examples are sufficient for training. Table 5 summarizes this list of interface design learnings including implemented features and improvements for future work.

Overall, the scope of this tool is limited to the use of bag-of-words as features for predicting binary concepts. A simple extension of the tool could support bigram, trigram and n-gram models. NLPReViz deals with a relatively simple machine learning problem of classifying documents. There is an opportunity to not only make predictions at a document-level

Table 5: **Summary of insights for the user interface design, grouped into the three steps of the learning cycle.**

Category	Learning
Review	1. Color schemes representing document classes and confidence levels helped the users easily interpret NLP results.
	2. Users found it easy to inspect the full-text of the documents with the important keywords highlighted. This worked well with the overview visualizations: the grid view and the WordTree views. This is in accordance to the popular visual information-seeking mantra of Overview, Zoom and Filter, and Details on demand [71].
	3. Users made little use of the overall dataset level statistics views including class distributions and overall list of important keywords used by the model. We hypothesize that this could be because of the fact that the participants were less familiar with the dataset beforehand. This may be more important for clinical care applications where the users may not be specifically interested in exploring the model per se but are instead primarily focused on patient care.
Feedback	1. Participants were able to intuitively use the three kinds of feedback mechanisms. Usage patterns differed between users. Some participants preferred the WordTree to provide feedback as it helped them label the documents more efficiently, others adopted more conservative ways to provide labels to reduce errors.
	2. Active learning approaches may be adopted to reduce the number of labels needed from the users.
	3. Another useful addition would be to address the difficulty in receiving evidence-based feedback for negative examples, ie. allowing the users to provide missing or absent evidence in a document, as a result of which it belongs to that class. Additionally, they requested support for indicating that specific phrases were irrelevant to a classification of a document. These can be used to reduce the weights against individual features contained in them.
Re-Train	1. Perform auto-retraining in the background when a sufficient number of feedback items have been provided, or by using other relevant heuristics. This would save the users from the need to click on retrain and create a new model manually.
	2. Users adopted ad-hoc methods, such as exploring the color spread of the WordTree to judge model quality. Support for built-in mechanisms to validate and generate model performance reports against a held-out test set could be a useful addition. This could continuously monitor model revisions and help the users understand their progress.

but also within different parts of the document as well. In the next prototype, as discussed in the following chapter, I build upon this work to train models for sentence-level predictions to identify important parts of the note to be included in a summary document. This is a step beyond NLPReViz and extends its goal from making predictions at document-level to identify relevant parts (or sentences) within a note. (Figure 9). Future enhancements might involve extending the interaction techniques to support feedback for other types of NLP models-such as extracting general concepts and relationships, summarization, question-answering, and so on. This opens the possibility of applying these strategies to a broad range of NLP problems (See Figure 4, in Chapter 1).

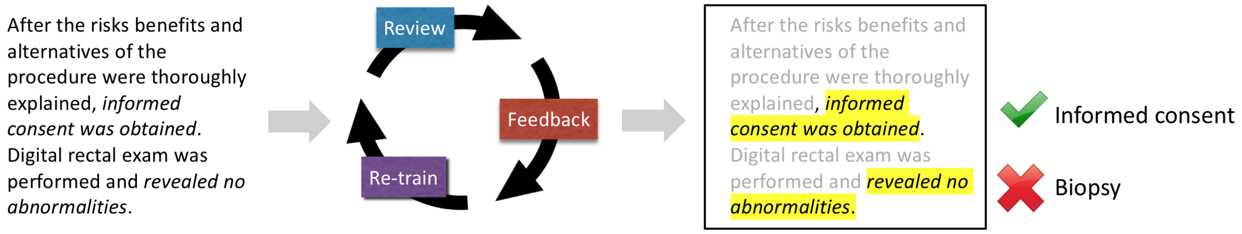

After the risks benefits and alternatives of the procedure were thoroughly explained, *informed consent was obtained*. Digital rectal exam was performed and *revealed no abnormalities*.
Review
Feedback
Re-train
After the risks benefits and alternatives of the procedure were thoroughly explained, *informed consent was obtained*. Digital rectal exam was performed and *revealed no abnormalities*.
✓ Informed consent
✗ Biopsy

Figure 9: **My new tool extends the learning task in NLPReViz (Chapter 4) from not only classifying documents to also identifying relevant text spans within them.**

Interactive Natural Language Processing has promise in both improving clinical care and further clinical research faster (Section 2.4.1). NLPReViz served as an example of how clinicians could train their own models for retrospective research. In the next Chapter, we will see how interactive NLP can be useful in a clinical care environment, where I build an interactive signout tool. To build this new tool, I will adopt the steps described in this chapter and start by laying out the design requirements. This makes it easier to ideate the design of the learning model and the interface components supporting the learning cycle. I will also conduct similar evaluation studies to evaluate usability and correctness. Figure 8 shows an example of how we can evaluate model performance as clinicians provide more input. All of these steps are discussed in the following chapter, where I described the design and evaluation of this new prototype.

5.0 *INTELLIGENT SIGNOUTS:* INTERACTIVE NLP IN CLINICAL CARE: IDENTIFYING INCIDENTAL FINDINGS FROM TRAUMA REPORTS

Free text clinical records are a key component of a patient's electronic medical record (EMR) that capture rich information about patients. Despite advances in natural language processing (NLP) techniques, extracting relevant information from free-text clinical records remains challenging and time-consuming [14]. Building user-centered tools that enable NLP to be interactive has the potential to make NLP more useful for clinical applications. Interactive tools that ease the construction, review and revision of NLP models can empower clinicians, administrators, and patients who are the end-users. Interactive Natural Language Processing has the potential to make NLP more useful in live clinical practice and also further clinical research faster. In the previous chapter, we explored its application for retrospective research on clinical notes. We need tools to ease the construction, review, and revision of Natural Language Processing models for applications in clinical care as well.

In this chapter, I present the design, implementation, and evaluation of an interactive NLP tool to help clinicians find relevant information in patient notes. This task is particularly relevant for the construction of summaries, which are often manually curated by clinicians for a variety of clinical tasks, including writing discharge summaries, preparing for rounding, and seeing new consults. To demonstrate the interactive learning approach in a clinical care setting, we will use our prototype tool to help trauma physicians identify incidental findings from radiology reports for preparing signout notes. Physicians and nurses use *signout* notes to provide concise summaries used to facilitate transitions in care between providers [72]. These signout notes summarize key observations and interpretations from a patient's medical record. Although electronic signout systems have been adopted in some healthcare organizations, construction of summaries is still a manual process. It is a repetitive task and carries a risk of errors and omissions. Interactive methods are particularly appealing here as they allow clinicians to train and revise their own models. Individual physicians and their teams have their own requirements and needs about how incidentals are defined and what should go into a signout note (eg. trauma vs. oncology). The definition of incidental findings is sensitive to the clinical context, e.g., the surgeon's notion of incidental findings in a trauma patient may be very different from an oncologists definition of incidental findings in a cancer patient. Such differences also arise during other tasks such as preparing history and physicals (H&Ps), consult notes, progress notes, and pre-rounding. This presents a challenge to automated extraction approaches based on limited training data, making the identification of incidental findings a task best served by models customized to the clinical group and context. Collecting customized training corpora for data-scientists to build models for different summarization tasks used by individual teams and institutions is not scalable. Interactive methods offer a feasible method to build a variety of customized models and address individual task needs. Overall, this prototype tool will serve as a demonstration of the interactive NLP approach. Lessons learned from the development and evaluation of this tool will provide insight into the generalized design of interactive NLP systems for wider clinical applications.

This new prototype extends the learning problem in NLPReViz [48] for identifying relevant text spans within a full-text patient note (See Table 1 in Chapter 2 for comparison). This can be seen as a step beyond the previous classification task – where we are not only classifying whether a document is important but are also interested in identifying relevant portions within the important document as well (Figure 9). It can be modeled as a binary classification problem to predict whether a particular sentence is 'important' or 'not important' to be included in the signout.

The interface components need a complete redesign yet fulfill the *review*, *feedback* and *retrain* steps of the interactive learning cycle. This is due to the differences in usage scenarios between the two applications. The users of NLPReViz were researchers who were vested in the task of building NLP models. In comparison to Intelligent Signouts, clinicians are primarily engaged in patient care and are building NLP models as a background task. In

Section 2.1, we discussed other examples of similar approaches for interactive learning for bug triaging [23], tailoring music and movie recommender systems [24] and even music composition [25].

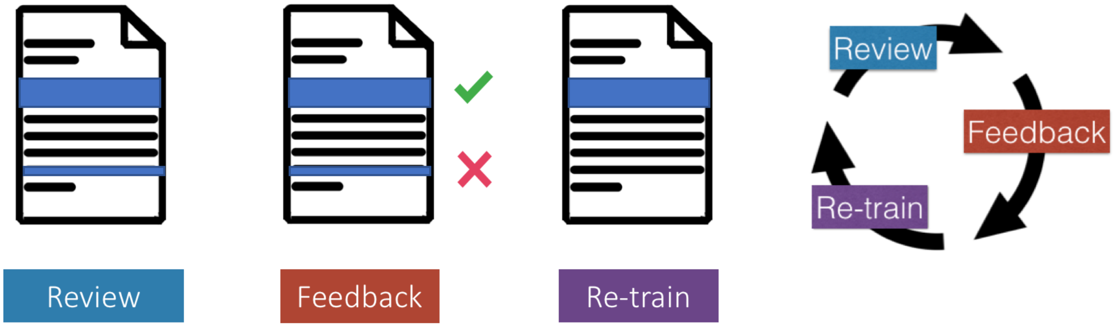


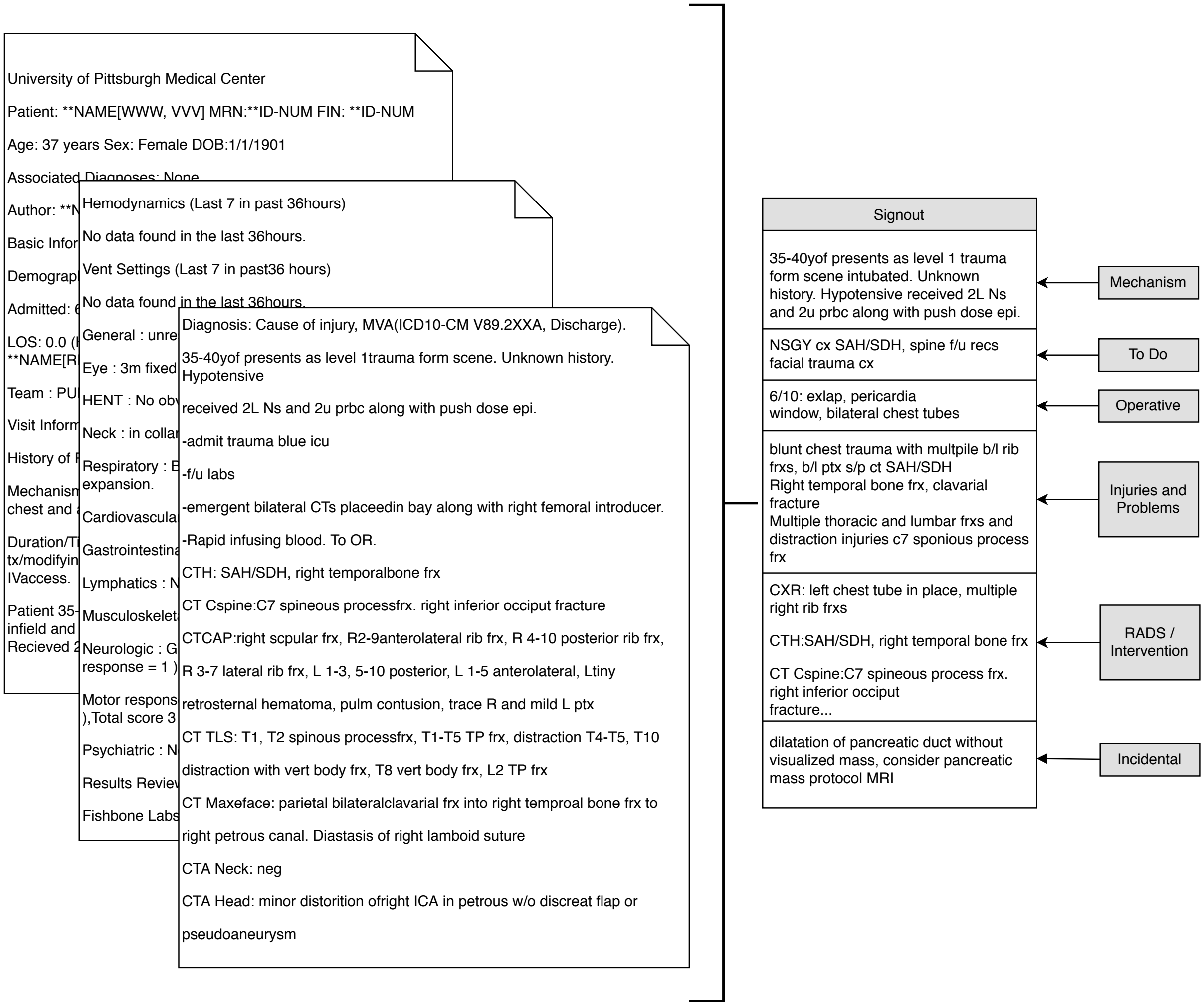
Figure 10: ***System overview:*** **Building iterative Learning Models for predicting important and relevant information within clinical notes. Physicians 1) review highlights predicted by the system, 2) and provide feedback on them. 3) Once these feedbacks are used to re-train models, it completes an interactive learning cycle.**

In this chapter, I present the design and implementation of a web-based tool interactive NLP tool for preparing signout notes followed by a user study with physicians to evaluate our prototype tool. Physicians may iteratively refine the NLP models by providing feedback (Figure 10). By extending this effort, we may envision an intelligent signout note creation system which can identify text and highlight them in the full-text report for inclusion in the signout notes.

5.1 SIGNOUT NOTES

Physicians and nurses use signouts to handoff patients from one team to another. It is used to transfer information about the patients under their care. Although traditionally this is done through verbal signout protocols, more recently hospitals are supporting structured, written signouts notes through custom-built applications [73, 74]. Written sign-out notes shared between physician teams allow for smooth transitions in care between shifts [72] and can serve as indicators of important and more relevant pieces of information in the patient's medical record. Despite the introduction of electronic signout systems, the process of abstracting and summarizing the full notes in the patients' EMR is done manually. Wohlauer et al. [73] conducted a study which revealed that residents spent 1-2 hours every day to gather important data from multiple sources. Further, there is scope for introducing errors when information is copied manually [11, 72, 75, 76]. When doing electronic sign-outs, physicians prepare the full-text report (often using dictation) and then write up the corresponding signout note for the patient at a later time. They must also be updated as more reports are filed in the EMR during a patient's stay. Although preparing and updating signout notes are time-consuming tasks, they are very useful to the team of physicians taking care of the patients. This problem is suitable for using our interactive NLP approach, where physicians could train NLP models to help them create signout notes, thereby addressing some of the problems with them.

Signout notes vary in structure between different teams. A trauma team may have very different guidelines about what should go into a signout as compared with say, clinicians in surgical oncology. Moreover, different institutions may impose different requirements and guidelines as well. These variations are tolerable as they are only used for internal communication. Figure 11 shows an example of a signout note from the University of Pittsburgh Medical Center's (UPMC) Trauma services. The note is divided often divided into different categories as well containing information about *Mechanism* - abstracting information in a Health and Physical Exam note, a *To Do* list, *Injuries and Problems*, *Radiology and Intervention* - summarizing reports from the radiologists, and *Incidentals* among other sections. At UPMC hospitals, physicians prepare signout notes using the Physician Sign-Out (PSO) Application. This is developed by the Custom Application Group at Information Services Division and resides outside of the patient's electronic medical record. The application manages and records signout for nearly 14,000 encounters every 3 months. These notes are also sometimes used for morning rounds apart from handoffs between shifts.


University of Pittsburgh Medical Center

Patient: **NAME[WWW, VVV] MRN:**ID-NUM FIN: **ID-NUM

Age: 37 years Sex: Female DOB:1/1/1901

Associated Diagnoses: None

Author: **N

Basic Infor

Demograp

Admitted: 6

LOS: 0.0 (
**NAME[R

Team : PU

Visit Inform

History of P

Mechanism
chest and a

Duration/Ti
tx/modifyin
IVaccess.

Patient 35-
infield and
Recieved 2

Hemodynamics (Last 7 in past 36hours)

No data found in the last 36hours.

Vent Settings (Last 7 in past36 hours)

No data found in the last 36hours.

General : unre

Eye : 3m fixed

HENT : No obv

Neck : in collar

Respiratory : B
expansion.

Cardiovascula

Gastrointestina

Lymphatics : N

Musculoskelet

Neurologic : G
response = 1)

Motor respons
),Total score 3

Psychiatric : N

Results Review

Fishbone Labs

Diagnosis: Cause of injury, MVA(ICD10-CM V89.2XXA, Discharge).

35-40yof presents as level 1trauma form scene. Unknown history. Hypotensive

received 2L Ns and 2u prbc along with push dose epi.

-admit trauma blue icu

-f/u labs

-emergent bilateral CTs placeedin bay along with right femoral introducer.

-Rapid infusing blood. To OR.

CTH: SAH/SDH, right temporalbone frx

CT Cspine:C7 spineous processfrx. right inferior occiput fracture

CTCAP:right scpular frx, R2-9anterolateral rib frx, R 4-10 posterior rib frx,

R 3-7 lateral rib frx, L 1-3, 5-10 posterior, L 1-5 anterolateral, Ltiny

retrosternal hematoma, pulm contusion, trace R and mild L ptx

CT TLS: T1, T2 spinous processfrx, T1-T5 TP frx, distraction T4-T5, T10

distraction with vert body frx, T8 vert body frx, L2 TP frx

CT Maxeface: parietal bilateralclavarial frx into right temproal bone frx to

right petrous canal. Diastasis of right lamboid suture

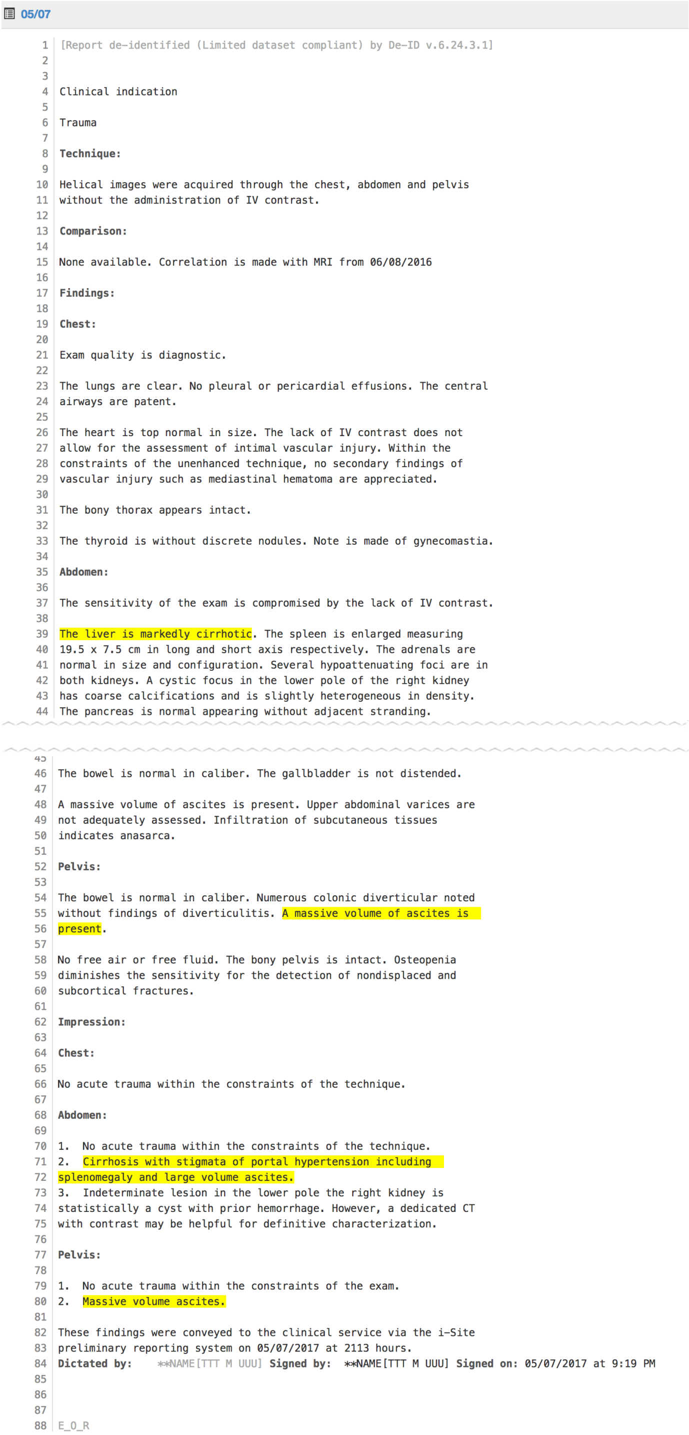
CTA Neck: neg

CTA Head: minor distorition ofright ICA in petrous w/o discreat flap or

pseudoaneurysm

Signout	
35-40yof presents as level 1 trauma form scene intubated. Unknown history. Hypotensive received 2L Ns and 2u prbc along with push dose epi.	Mechanism
NSGY cx SAH/SDH, spine f/u recs facial trauma cx	To Do
6/10: exlap, pericardia window, bilateral chest tubes	Operative
blunt chest trauma with multpile b/l rib frxs, b/l ptx s/p ct SAH/SDH Right temporal bone frx, clavarial fracture Multiple thoracic and lumbar frxs and distraction injuries c7 sponious process frx	Injuries and Problems
CXR: left chest tube in place, multiple right rib frxs CTH:SAH/SDH, right temporal bone frx CT Cspine:C7 spineous process frx. right inferior occiput fracture...	RADS / Intervention
dilatation of pancreatic duct without visualized mass, consider pancreatic mass protocol MRI	Incidental

Figure 11: **Example of a signout note compared with a history and physical examination (H&P) note. Signout notes often summarize information from H&P, Progress notes, Radiology reports and also Operative notes. Incidentals is one of the main sections of this note.**


05/07

[Report de-identified (Limited dataset compliant) by De-ID v.6.24.3.1]

Clinical indication

Trauma

Technique:

Helical images were acquired through the chest, abdomen and pelvis without the administration of IV contrast.

Comparison:

None available. Correlation is made with MRI from 06/08/2016

Findings:

Chest:

Exam quality is diagnostic.

The lungs are clear. No pleural or pericardial effusions. The central airways are patent.

The heart is top normal in size. The lack of IV contrast does not allow for the assessment of intimal vascular injury. Within the constraints of the unenhanced technique, no secondary findings of vascular injury such as mediastinal hematoma are appreciated.

The bony thorax appears intact.

The thyroid is without discrete nodules. Note is made of gynecomastia.

Abdomen:

The sensitivity of the exam is compromised by the lack of IV contrast.

The liver is markedly cirrhotic. The spleen is enlarged measuring 19.5 x 7.5 cm in long and short axis respectively. The adrenals are normal in size and configuration. Several hypoattenuating foci are in both kidneys. A cystic focus in the lower pole of the right kidney has coarse calcifications and is slightly heterogeneous in density. The pancreas is normal appearing without adjacent stranding.

The bowel is normal in caliber. The gallbladder is not distended.

A massive volume of ascites is present. Upper abdominal varices are not adequately assessed. Infiltration of subcutaneous tissues indicates anasarca.

Pelvis:

The bowel is normal in caliber. Numerous colonic diverticular noted without findings of diverticulitis. A massive volume of ascites is present.

No free air or free fluid. The bony pelvis is intact. Osteopenia diminishes the sensitivity for the detection of nondisplaced and subcortical fractures.

Impression:

Chest:

No acute trauma within the constraints of the technique.

Abdomen:

1. No acute trauma within the constraints of the technique.
2. Cirrhosis with stigmata of portal hypertension including splenomegaly and large volume ascites.
3. Indeterminate lesion in the lower pole the right kidney is statistically a cyst with prior hemorrhage. However, a dedicated CT with contrast may be helpful for definitive characterization.

Pelvis:

1. No acute trauma within the constraints of the exam.
2. Massive volume ascites.

These findings were conveyed to the clinical service via the i-Site preliminary reporting system on 05/07/2017 at 2113 hours.
Dictated by: **NAME[TTT M UUU] **Signed by:** **NAME[TTT M UUU] **Signed on:** 05/07/2017 at 9:19 PM

E_O_R

Figure 12: **Example of radiology note with incidental findings highlighted. The patient was admitted into trauma after a fall from a step stool. The CT scan reveals massive volume *ascites* and *cirrhotic* changes as incidental findings. These findings are also repeated in the 'Impression' section. This is one of the 6 radiology reports for this patient.**

5.1.1 Identifying Incidental findings for preparing signouts

Within signout notes, I focused on the *'incidentals'* section for my studies to define tractable user-study tasks and build gold-standard datasets. The modern care of trauma patients relies on extensive use of whole-body computed tomography (CT) imaging for assessment of injuries [77]. While the CT imaging is invaluable in demonstrating the extent of the injuries, additional incidental findings are often uncovered such as occult masses, lesions, and anatomic anomalies, that are unrelated to the trauma [78]. For example, CT imaging in a person who had a fall may reveal a nodule that is unrelated to the injuries caused by the fall (see Figure 12). Incidental findings range from an insignificant renal cyst to a serious lung nodule. [79]. The members of the trauma team are responsible for interpreting the radiology reports, identifying and assessing the incidental findings, and conveying this information to the patient and other physicians. However, in a busy trauma center, this task can tax the team that is responsible for evaluating and treating the more pressing acute injuries [76]. A tool that can automatically identify and highlight relevant incidental findings would be an invaluable aid to the trauma team.

The current workflow for preparing the signout notes at the Trauma Services at UPMC is a manual process. It requires physicians to navigate between two different software systems. First, they go through the full-text notes from the patient's EMR (Cerner), synthesize them and then switch to the Physician Sign-Out application to fill in different sections of a templated signout note. This process is repeated and the signout note is revised whenever a new (full-text) note is added to the patient's EMR. Typically, resident physicians in the trauma team that include surgery, internal medicine and radiology residents are responsible for writing the signout notes. 'Incidentals' is one of the seven main sections in a signout note at UPMC Trauma service. I conducted preliminary meetings with the PSO Application developers to understand the underlying data format and structure of these notes in the dataset we received. I also completed an informal shadowing session in the trauma ICU to observe physicians doing signout. These sessions were conducted not with the goal of collecting formal data for research but serve as an initial validation of the problem and requirements. I used these sessions to gather initial insights and informal feedback towards developing the ideas in this work.

Restricting the scope of the problem to incidentals allows me to run evaluations with a reasonable amount of annotation effort towards creating a gold-standard set. This also helps me demonstrate the interactive learning approach with clear goals for the user studies participants, who may otherwise not be familiar with the complete guidelines for preparing an entire signout note at UPMC Trauma Services. Besides this, the problem of documenting incidentals is also interesting to the medical community as the existing system requires significant manual effort and poses risks of errors and omissions [75, 76]. Drawing comparisons with NLPReViz, this is similar to restricting the user study to two variables in [48] so that participants could devote a reasonable amount of time to train models.

There is little prior work in informatics research that use NLP models to help physicians in creating signouts. Yetisgen et al. [80] demonstrated the use of NLP and supervised machine learning for identifying critical recommendation sentences in radiology reports. They follow an extractive summarization approach and define their problem as a binary classification of sentences to extract critical findings. They start by building a corpus of 800 manually annotated radiology reports in order to train their learning models. In their work, they achieved 95.60% precision, 79.82% recall, 87.0% F-score, and 99.59% classification accuracy (with 5-fold cross-validation) in identifying the critical recommendation sentences in radiology reports [80]. Zech et al. also present more recent work on identifying findings from radiology reports demonstrates a similar pipeline and explores linear classification models [81]. It also notes that simple bag-of-words (unigram) methods performed competitively with other sophisticated methods tried in identifying findings. Yetisgen et al. also conducted a follow-up study to build a large annotated training corpus [82]. They noted that because manual annotation is a time-consuming and labor-intensive process, they could annotate only a small portion of their corpus. They recommend an interactive approach for creating such annotated corpora in their work. To the best of my knowledge, no prior work seeks to address the problem of collecting training data to build such models. The traditional machine approach would be to create annotated sets of training data in batches and is prone to the issues discussed earlier in Chapter 2. Using the interactive learning approach, we are able to build a continuously learning intelligent system which can revise NLP models as it learns from physicians' use over time (Figure 13).

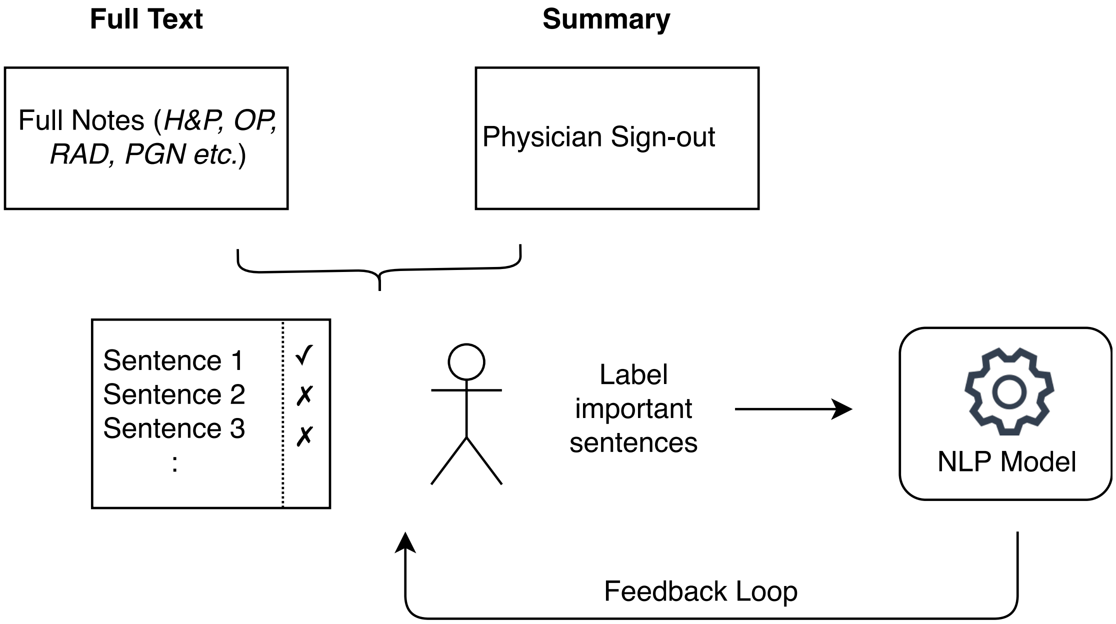


Figure 13: **Intelligent Signout Tool that learns from physicians' use over time.**

5.1.2 Modeling as an information extraction task

The problem of identifying relevant or important parts of a clinical report may also be also defined as an extractive text summarization task. In NLP literature, summaries are typically divided into *extractive* and *abstractive* summarization [83]. Extractive summaries contain material taken directly from the original documents, while abstracts synthesize a material that may not be present in the original form. Another classification for summarization is based upon how the summarization output is presented: *Indicative* summaries highlight important pieces in the original text, and Informative summaries are designed to replace the original text. In my work, I restrict my scope on *extractive* and *indicative* summarization, which is suitable for the problem of identifying sentences to help physicians prepare signout notes.

Extractive summarizers identify the most important sentences in a document or a group of documents. In their survey of text summarization techniques, Nenkova and McKeown

[84] identify that most extractive summarizers perform three tasks:

1. *Derive intermediate representation:* First, we derive some intermediate representation of the text to be summarized. We can use topic representation techniques such as frequency, TF-IDF, topic word and so on for this purpose. We can also generate additional features for machine learning by transforming each sentence as a list of indicators – sentence length, the presence of certain phrases, etc.
2. *Score sentences:* Next, each sentence is assigned a score indicating its importance. In machine learning methods, the score of each sentence is determined by the weights for different indicators defined in the intermediate representation.
3. *Select summary sentences:* Finally, we select the best combination of sentences for the summary. This can be done in different ways, selecting n-best scores, iterative greedy procedures, a global selection of a group of sentences among others.

Machine learning methods for extractive summarization use a representation of the text that can be used to score the sentences. When using supervised methods, this task can be framed as a binary classification problem – each sentence can either belong to a summary or non-summary class. The classification function scores each sentence based on the intermediate representation as inputs. Some common features as discussed above include the position of the sentence in the document, sentence length, similarity with title or headings, presences of cue phrases, the presence of named entities, etc. The task of the classifier is to estimate a probability score that a sentence will be in the summary, given the features present in it [85]:

$$P(s \in S | F_1, F_2, ..., F_k)$$

Most existing supervised learning algorithms are applicable for this task. However, we need an annotated datasets for training them. Interactive Machine Learning Systems are appealing for building models for such NLP tasks, which require expert constructed training data and examples. Using traditional approaches, the models are built by experts in linguistics and/or machine learning, which restricts the end-users to tweak them. There has been some prior work on summarizing clinical text using interactive tools with partial end-user involvement. However, an end-to-end interactive learning approach for text summarization would involve building revised models based on user feedback. There has been little work done towards solving the barriers to NLP which is in collecting relevant training data for building these models.

5.2 OBJECTIVES

5.2.1 Design and implement a prototype interactive tool

In this section, I present the design requirements for building an intelligent signout tool to be used by trauma care team for identifying incidental findings. We consider identifying incidentals as an example use-case for interactive NLP systems in clinical care. I built on my previous work on NLPReViz to address the challenge of integrating interactive NLP into the clinical workflow. The tool consists of 1) a user interface that enables users review, provide feedback and understand changes to the NLP model, and 2) a learning pipeline that builds, applies and updates an NLP model for identifying incidental findings.

5.2.1.1 User Interface The proposed tool is targeted towards clinicians who are primarily engaged in the care of individual patients. The task of identifying incidentals is a background task as they read the patient notes. In comparison, the users of NLPReViz were primarily interested in building NLP models in Chapter 4. Thus, the interface components need to be redesigned to fulfill the *review*, *feedback* and *retrain* steps of the interactive learning cycle (Table 1). Moreover, NLPReViz focused on making predictions at a document-level for clinical research. In this project, we aim to describe an example use-case scenario in clinical care for identifying sentences or phrases within a note or a document. Further, the users of NLPReViz were solely vested on the task of building the models, while in the new tool model-building moves to the background as the clinicians work in preparing signout notes (See Table 1). The design requirements for the new tool overlap with those discussed in the literature review of interactive machine learning in Section 2.1, and also NLPReViz. Using the ideas described in these prior works as suggestions for design, I present a shortlist some of the design ideas that are applicable to the new tool.

Design Requirements

The user interface should have functionality to help physicians in selecting relevant training examples and in providing labels appropriate for updating the NLP model. The interface should *display predictions* from the model and allow physicians to *give feedback* that will be used to *revise* the model. Visualization and interaction components should support these steps within the interactive learning cycle. These requirements are further itemized as follows:

(i) **Review**

R1: The user interface should highlight sentences as predicted by the NLP model to be relevant and, where possible, help users understand why a sentence was predicted to describe an incidental finding.

R2: The interface should help users to quickly navigate between documents as well as predictions.

(ii) **Feedback**

R3: Users should be able to select sentences that should have been highlighted and were missed by the NLP model. Similarly, they should be able to remove incorrect highlights.

R4: The user interface should help minimize user actions and time required for providing feedback.

(iii) **Re-train**

R5: Feedback provided by users should be displayed as a list of additions and deletions to help users understand changes between model revisions.

5.2.1.2 Learning Pipeline Requirements Our system needs a learning pipeline for making predictions about a clinical note. This involves pre-processing step including breaking up of patient notes into meaningful chunks, such as reports, sections, and sentences. Machine learning-based methods for sectionizing clinical notes include using models such as Bayesian Scoring [86] and Hidden Markov Models [87], etc. Simpler approaches include both rule-based and statistical models for sentence boundary detection.

Next, we need a feature extraction step which feed into a classification model, such as [80]. The learning problem may be modeled as a binary classification task of predicting whether text elements (sections or sentences) discuss relevant incidental findings or not. Figure 14 shows an example text processing pipeline (non-interactive) to identify critical recommendation sentences in radiology reports using a similar learning model used in [80]. They also experimented with other linguistic features such as whether the span included a modal verb, temporal phrase, etc. The results from Yetisgen et al. [80] and Zech et al. [81] (See Section 5.1.1), suggest that a similar model may be good enough for my Intelligent Signouts tool as well. In these projects, simple bag-of-words methods performed competitively with other sophisticated methods for classifying relevant sentences in radiology reports.

The learning problem can thus be modeled as a binary classification task of predicting whether a particular sentence is 'important' or 'not important' to be discussed in the signout. This can be seen as a step beyond the classification task seen in NLPReViz, where we are not only classifying whether a document is important but are also interested in identifying relevant portions within the important document as well. The system will incorporate user feedback to improve the learning model by processing the user input to make revisions. A simple implementation would be to use the 'rationale' based SVM models to build these revisions as in NLPReViz [64]. This would complete an interactive learning model that can be used for prediction useful elements from the full-text reports for preparing signout notes. More sophisticated classification approaches such as learning from only positive and unlabeled data [88], and more recent neural network-based approaches, such as [89, 90, 91], is beyond the scope of this work. The simpler pipelines providing competitive performance can be quickly implemented for demonstrated our interactive approach. Future work may extend these methods to reflect state-of-the-art NLP methods.

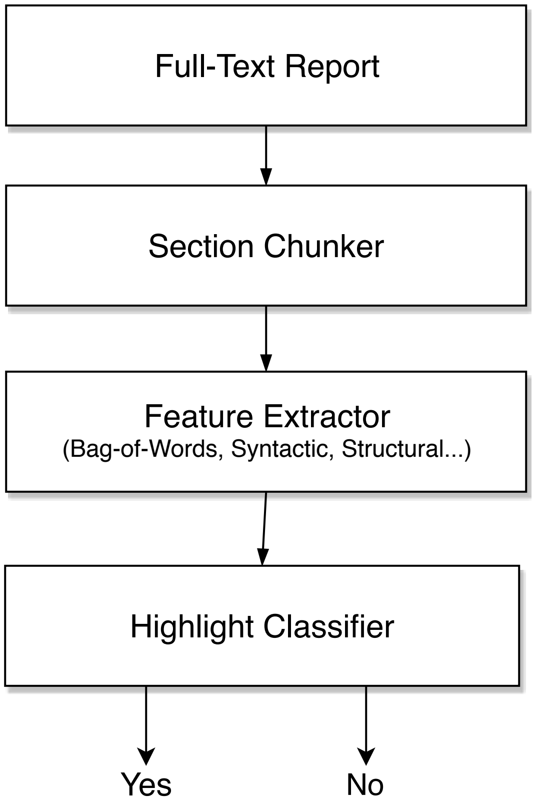


Figure 14: **Example of a classification pipeline to identify relevant sentences used by Yetisgen et al.** [80]

5.2.2 Hypotheses

We hypothesize that our tool will enable physicians to build useful NLP models for identifying incidental findings in radiology reports within a closed feedback loop, with no support from NLP experts. We may further split this into two sub-hypotheses for usability and correctness:

H1: *The interactive tool will be used by physicians successfully to identify incidental findings with little or no support from NLP experts.*

Design of interactive learning systems require that we adopt a human-centered approach for collecting training data and building models. Simple active learning approaches that involve asking a series of questions to human oracles" can be annoying and frustrating, as noted in Section 2.1. The focus in IML is in building tools that align the process of providing feedback with user needs. Thus, we test whether the proposed tool is usable by end-users, i.e., physicians. for the task of identifying incidental findings.

H2: *The interactive tool will decrease time and effort for physicians for identifying incidental findings.* Interactive machine learning systems are designed to support an iterative learning cycle instead of asking the users to work in long batches seen in traditional methods. But these models need to be evaluated for correctness and usefulness. This allows us to demonstrate the value of building the models interactively and justify the costs involved in doing so. In my work, I aim to evaluate both the time and effort as well as how the interactive cycle could help building useful models that can be revised to reach high levels of accuracy. We compare our IML approach to a simpler interface lacking IML, using measurements of time and effort (in terms of number of user actions) to evaluate how the interactive cycle could facilitate construction of highly-accurate models.

5.3 METHODS AND MATERIALS

We followed a three-step sequence for design, implementation, and evaluation for our tool. We began by discovering and defining the design requirements for the system. The system comprises of a) the learning model that makes the predictions and makes sense of user feedback to revise the models and b) the visualization and interaction methods to support this feedback loop. For building the prototype, we followed an iterative process starting with design mock-ups, followed by implementation and revision phases. Both the learning model and interface components were tested for bugs and issues that may otherwise interfere with the evaluation. We also created a labeled gold standard dataset for running our evaluation studies. These steps are described in detail in the following sections.

5.3.1 Dataset

We obtained a 3-month long snapshot of de-identified Physician Signout dataset from UPMC Presbyterian's Trauma Service. It consists of both the signout notes along with corresponding full-text notes (History and Physical, Progress Notes, Operative Notes, and Radiology Reports). This dataset was obtained using IRB PRO17030447. It comprises of 75,946 signout notes (including revisions) and 192,347 full-text notes in the EMR. The average length of the full-text reports is 6,000 characters each with a patient having an average of 18 notes (1 to 431, median = 10), as compared to 1,000 characters in a signout note. Thus, signout notes are nearly 108 times shorter than the full-text notes for a patient on average.

To create an annotated dataset, two trauma physicians annotated 4,181 radiology reports (686 encounters, 6.09±4.18 reports per encounter following a power-law distribution) for incidental findings using a custom annotation tool. Annotators focused on two types of incidental findings that are recommended for follow-up: lesions suspected to be malignant and arterial aneurysms meeting specified size and location criteria. Table 6 provides detailed annotation guidelines that were used by the physicians. An initial pilot set of 128 radiology reports was annotated by the two physicians independently, and the inter-annotator agreement (IAA) measured using Cohen's Kappa statistic [70] was 0.73. After review and deliberation, the annotation guidelines were revised, and a second pilot set of 144 radiology reports was annotated. This resulted in a revised IAA of 0.83. Each of the remaining 4,053 reports was annotated by a single physician using the revised annotation scheme.

We sampled a subset of encounters from this annotated dataset for our evaluation user study as described later in Section 5.3.5. We restricted the sample to only those encounters with at least one or more incidentals findings. Further, we considered only those encounters which had between 3-7 reports per case. This allowed us to avoid outliers with large numbers of reports to allow for a reasonably consistent review time duration per encounter. Annotators (same physicians) reviewed this smaller sample of 694 reports (130 encounters; 5.36±1.3 reports per encounter; mostly CT and X-ray reports, with a small number of other modalities such as ultrasound, magnetic resonance imaging, fluoroscopy, etc.) again to remove any inconsistencies in labeled gold standard against the annotation guidelines (Table 6). This sample with revised annotations was used in the user study.

5.3.2 Learning Pipeline

We extracted individual sentences using spaCy (a Python NLP library: https://spacy.io [92]). A sentence was labeled positive if a phrase in it or the entire sentence was selected by the annotators. Sections were extracted after applying regular expressions to identify section headings. Similarly, a section was marked positive if it contained one or more sentences with incidental findings. Table 7 shows the distribution of incidentals over these levels.

We used a simple NLP pipeline using a bag-of-words feature-set along with a classifier using support vector machines with a linear kernel. Earlier results suggest that this approach performed competitively with other sophisticated methods for classifying relevant sentences in radiology reports [80, 81]. We used the 'rationale model' proposed by Zaidan et al. [64] for implementing interactive machine learning with user feedback. Specifically, when the user identified a span of text as an incidental finding, we constructed similar synthetic text as additional training data. Using a simple classification model allowed us to focus the discussion in this paper on the design of the overall system. We performed a detailed exploration into classifier modeling techniques for identifying incidental findings is out of the

Table 6: ***Annotation guidelines:*** **Adapted from Sperry et al. [76]. Any lesion of malignant potential and any arterial aneurysm that is greater than a specified size was annotated.**

Lesions	
Brain	Any solid lesion
Thyroid	Any lesion
Bone	Any osteolytic or osteoblastic lesion, not age-related
Breast	Any solid lesion
Lung	Any lesion
Liver	Any heterogeneous lesion
Kidney	Any heterogeneous lesion
Adrenal	Any lesion
Pancreas	Any lesion
Ovary	Any heterogeneous lesion
Bladder	Any lesion
Prostate	Any lesion
Intraperitoneal/ Retroperitoneal	Any free lesion

Aneurysms	
Thoracic aorta	$\geq$ 5cm
Abdominal aorta	$\geq$ 4cm
External iliac artery	$\geq$ 3cm
Common femoral artery	$\geq$ 2cm
Popliteal artery	$\geq$ 1cm

Table 7: **Prevalence of incidental findings in the data sampled from the annotated dataset. Positives denote the raw count of sentences, sections, or reports containing one or more incidental findings.**

	Total	Positives	Prevalence
Reports	694	164	23.63%
Sections	6046	302	5%
Sentences	20,738	369	1.78%

scope of this dissertation, but is discussed in another manuscript [93].

5.3.3 Interface Design

Tables 2 and 3 list several ideas used in prior work in interactive machine learning for addressing the design requirements described above. These tools are described in detail in Chapter 1, demonstrate the following design suggestions that may be relevant for the prototype tool:

(i) **Review:** Most tools such as EluciDebug [37], UTOPIAN [59] and NLPReViz [48] provide good examples for interface views to display NLP classification results using both in-place highlights and statistics views showing feature weights. CueFlik [38, 47] and Apolo [26] provide examples of views to help the users visualize examples from each classification. Other approaches demonstrated in these works is the use of data-set level visualizations such as word-clouds to visualize documents in each class.

Visual Classifier Training [27] presents an interactive view of the classifier that helps the users interpret the decision boundary and prediction confidence. This view may not be very relevant for our work as the users are not solely vested in the task of building NLP models. In NLPReViz, we made use of color saturation and mouseover text to convey the confidence levels to the users. Users also had an option to sort the documents based on confidence levels and spend more effort on low-confidence predictions.

Overall, CueT [23] shows an example of how interactive machine learning may be built as a background task to support the user's primary objective, i.e. triage bugs. It demonstrates how the review step can be designed without disrupting the primary task that the users are engaged in (which is finding key sentences, in the proposed prototype) and have interactive machine learning augment it. While doing so, we would also need to support navigation and views for review following the information-seeking mantra of Overview, Zoom and Filter, and Details on demand [71].

(ii) **Feedback:** NLPReViz [48] explores three different ways for the users to provide feedback: 1) document-level labels, 2) feedback using rationales, and 3) WordTree view. It demonstrates how the feedback step can be made more efficient using the WordTree view. It was useful for browsing through several patients together at once. It visualizes common phrase structures across different notes which is relevant for retrospective research. However, in the prototype tool, although the system learns from all the patients' records, users are interested in browsing one patient at a time for preparing a signout note. Thus, such views may not be very useful for the initial prototype.

EluciDebug [37], Movie Tuner [24] and Visual Classifier Training [27] implement views for the users to manipulate feature weights at both document and dataset levels. Some of the techniques to provide feedback features from individual example from the report currently being reviews would be useful for the prototype tool.

Teaching Simon [39] notes that teacher-triggered feedback, as opposed to a continuous stream of questions from the system resulted in better satisfaction from the users while learning as fast as active learning approaches. Similarly, Interactive medical word sense disambiguation [50] and others also suggested that an interactive approach where the users can provide richer feedback [43] outperformed traditional active learning methods with limited scope for users to provide input. Thus, rationale-based approaches used in NLPReViz are suitable for adoption in the new problem as well.

NLPReViz provides examples of views for resolving conflicts and errors. Otherwise, prior work pays little attention to the problem of users introducing annotation errors during feedback.

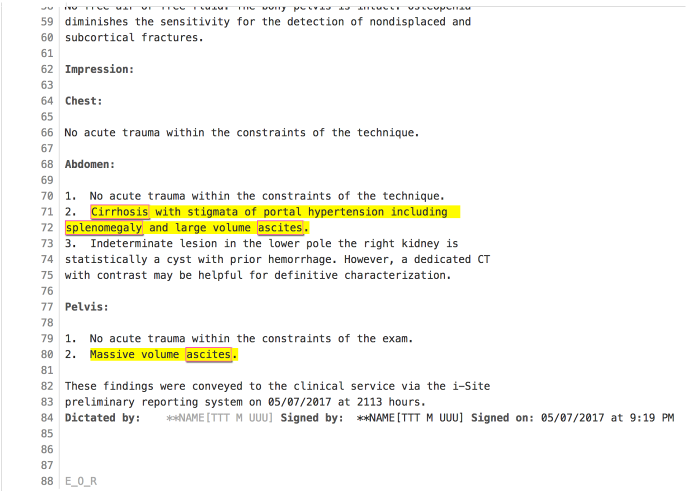
(iii) **Retrain:** Tools such as UTOPIAN [59] support views for showing revisions between model iterations. In NLPReViz, users were able to see changes to the model based on their feedback as the tool indicated which documents switched labels. EluciDebug [37] and Visual Classifier Training [27] also visualize changes in feature weights at a dataset level as the models are revised. The dataset level views may be a lower priority for the new prototype due to the same reasons as discussed before.

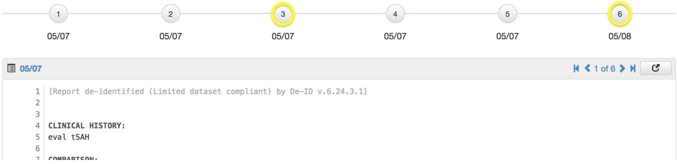
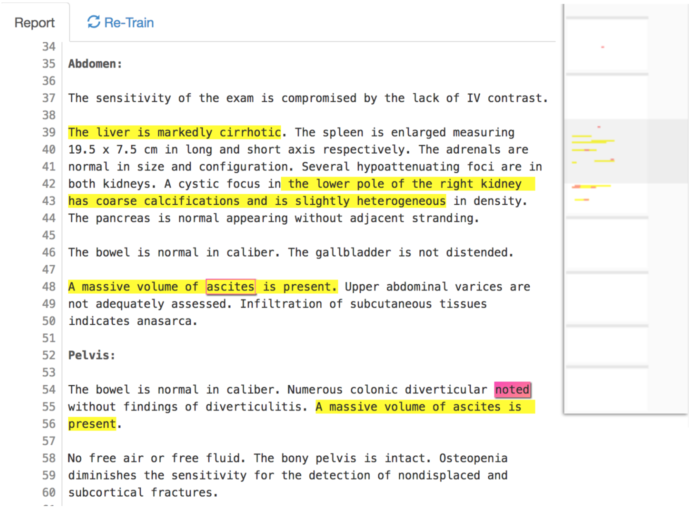
NLPReViz allowed the users to view annotation progress with a progress bar describing the number of examples labeled. But we did little to address the problem of letting the users know when the model is good enough. A visualization of model performance metric over time could help address this need. This could be done based on how the model performs on a held-out set and/or the set of training examples already reviewed and labeled manually. Another feature demonstrated in EluciDebug [37] and UTOPIAN [59] to allow users revert changes and switch between model revisions.

Some other enhancements could include helping users tune the parameters of learning model, such as in ManiMatrix [41] to decide model tradeoffs, or expose more of the steps involved in the learning model, such as those seen in toolkits providing GUI support to data-scientists [57, 56, 12].

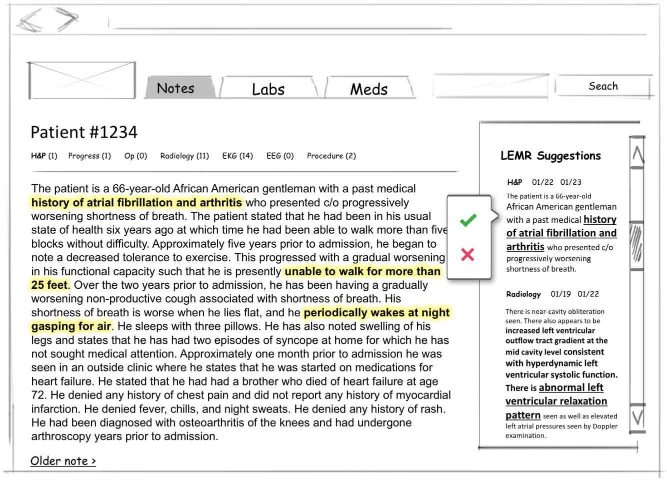
Many of these features may be considered as incremental enhancements to the prototype. Based on the ideas discussed above, I present a list of prioritized features to be implemented for evaluating the interactive learning approach in Table 8. They cover different visual interface views and interactions for the three components of the interactive learning cycle:

Table 8: **List of prioritized solutions to address the design requirements of the signout tool. They are grouped the three steps of the learning cycle: review, feedback & retrain.**

Step	Requirement	Description
Review	R1: Visual displays should highlight important sentences as predicted by the NLP model and, where possible, help the users understand why an incidental was predicted.	A simple way of marking important sentences would be through the use of color, such as using yellow highlights. A lighter shade represented lower confidence, while high confidence predictions were marked by a darker background color. Font size could also be used here. Confidence percentages may also be shown upon mouse over. We may use feature highlighting to show important features that lead to the prediction of relevant sentences. Users may also choose to reveal features for non-relevant sentences helping the, understand why they were not highlighted.  No free air or free fluid. The bony pelvis is intact. Osteopenia diminishes the sensitivity for the detection of nondisplaced and subcortical fractures. **Impression:** **Chest:** No acute trauma within the constraints of the technique. **Abdomen:** 1. No acute trauma within the constraints of the technique. 2. Cirrhosis with stigmata of portal hypertension including splenomegaly and large volume ascites. 3. Indeterminate lesion in the lower pole the right kidney is statistically a cyst with prior hemorrhage. However, a dedicated CT with contrast may be helpful for definitive characterization. **Pelvis:** 1. No acute trauma within the constraints of the exam. 2. Massive volume ascites. These findings were conveyed to the clinical service via the i-Site preliminary reporting system on 05/07/2017 at 2113 hours. **Dictated by:** **NAME[TTT M UUU] **Signed by:** **NAME[TTT M UUU] **Signed on:** 05/07/2017 at 9:19 PM E_O_R

	R2: Interaction components should help users quickly navigate between documents as well as predictions.	A design mockup for navigating between notes in chronological order marking those that contain highlights is shown below. Clicking on specific date events scrolls to that note.  1 2 3 4 5 6 05/07 05/07 05/07 05/07 05/07 05/08 05/07 1 of 6 [Report de-identified (Limited dataset compliant) by De-ID v.6.24.3.1] CLINICAL HISTORY: eval tSAH COMPARISON: For long notes, we could also make use of a page map showing positions of the highlights on the page along with the current view position of the note. This would allow users to quickly jump to interesting sections.  **Abdomen:** The sensitivity of the exam is compromised by the lack of IV contrast. The liver is markedly cirrhotic. The spleen is enlarged measuring 19.5 x 7.5 cm in long and short axis respectively. The adrenals are normal in size and configuration. Several hypoattenuating foci are in both kidneys. A cystic focus in the lower pole of the right kidney has coarse calcifications and is slightly heterogeneous in density. The pancreas is normal appearing without adjacent stranding. The bowel is normal in caliber. The gallbladder is not distended. A massive volume of ascites is present. Upper abdominal varices are not adequately assessed. Infiltration of subcutaneous tissues indicates anasarca. **Pelvis:** The bowel is normal in caliber. Numerous colonic diverticular noted without findings of diverticulitis. A massive volume of ascites is present. No free air or free fluid. The bony pelvis is intact. Osteopenia diminishes the sensitivity for the detection of nondisplaced and subcortical fractures.

To help navigate between notes, we could also use a suggestions box (shown on the right, in the following image) containing phrases from the full-text report identified to be included in the signout note. It also allows the users to quickly navigate between different suggestions made by the system.



Patient #1234

H&P (1) Progress (1) Op (0) Radiology (11) EKG (14) EEG (0) Procedure (2)

The patient is a 66-year-old African American gentleman with a past medical **history of atrial fibrillation and arthritis** who presented c/o progressively worsening shortness of breath. The patient stated that he had been in his usual state of health six years ago at which time he had been able to walk more than five blocks without difficulty. Approximately five years prior to admission, he began to note a decreased tolerance to exercise. This progressed with a gradual worsening in his functional capacity such that he is presently **unable to walk for more than 25 feet**. Over the two years prior to admission, he has been having a gradually worsening non-productive cough associated with shortness of breath. His shortness of breath is worse when he lies flat, and he **periodically wakes at night gasping for air**. He sleeps with three pillows. He has also noted swelling of his legs and states that he has had two episodes of syncope at home for which he has not sought medical attention. Approximately one month prior to admission he was seen in an outside clinic where he states that he was started on medications for heart failure. He stated that he had had a brother who died of heart failure at age 72. He denied any history of chest pain and did not report any history of myocardial infarction. He denied fever, chills, and night sweats. He denied any history of rash. He had been diagnosed with osteoarthritis of the knees and had undergone arthroscopy years prior to admission.

Older note >

✓ ✗

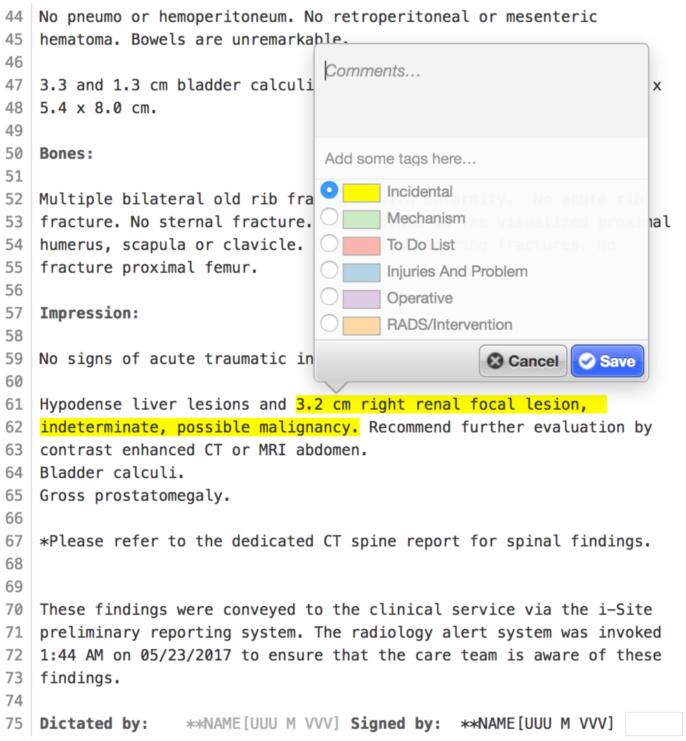
LEMR Suggestions

H&P 01/22 01/23

The patient is a 66-year-old African American gentleman with a past medical **history of atrial fibrillation and arthritis** who presented c/o progressively worsening shortness of breath.

Radiology 01/19 01/22

There is near-cavity obliteration seen. There also appears to be **increased left ventricular outflow tract gradient at the mid cavity level consistent with hyperdynamic left ventricular systolic function. There is abnormal left ventricular relaxation pattern** seen as well as elevated left atrial pressures seen by Doppler examination.

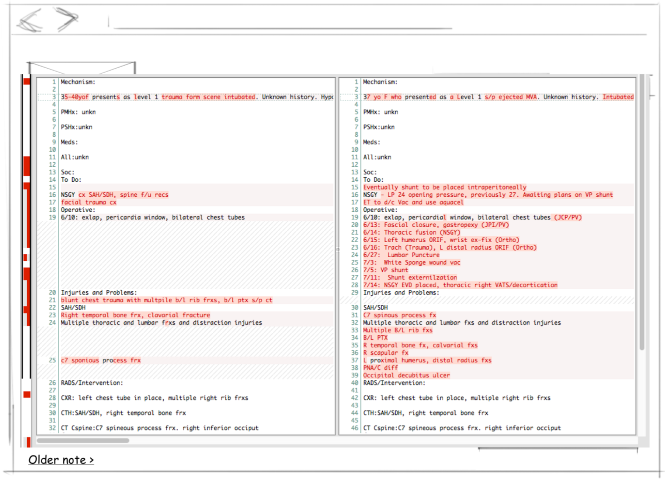
Feedback	R3: Users should be able to select sentences that must be highlighted and were missed by the NLP model. Similarly, they should be able to remove non-useful highlights.	Physicians could provide feedback by selecting and deselecting important portions for text as explicit feedback for the learning system. This can be part of their regular workflow while preparing signout notes. The interface could also help them categorize these highlights into different sections in a signout note. Navigation views, such as a suggestion box may also be used to quickly confirm or remove feedback. Different affordances may be implemented to cater to this need using mouseover, clicks, etc. 

```
No pneumo or hemoperitoneum. No retroperitoneal or mesenteric
hematoma. Bowels are unremarkable.

3.3 and 1.3 cm bladder calculi                                   x
5.4 x 8.0 cm.

Bones:

Multiple bilateral old rib fra
fracture. No sternal fracture.                                  al
humerus, scapula or clavicle.
fracture proximal femur.

Impression:

No signs of acute traumatic in

Hypodense liver lesions and 3.2 cm right renal focal lesion,
indeterminate, possible malignancy. Recommend further evaluation by
contrast enhanced CT or MRI abdomen.
Bladder calculi.
Gross prostatomegaly.

*Please refer to the dedicated CT spine report for spinal findings.


These findings were conveyed to the clinical service via the i-Site
preliminary reporting system. The radiology alert system was invoked
1:44 AM on 05/23/2017 to ensure that the care team is aware of these
findings.

Dictated by:    **NAME[UUU M VVV] Signed by:   **NAME[UUU M VVV]
```

Comments...

Add some tags here...

● Incidental
○ Mechanism
○ To Do List
○ Injuries And Problem
○ Operative
○ RADS/Intervention

	R4: Users should be able to select sentences that must be highlighted and were missed by the NLP model. Similarly, they should be able to remove non-useful highlights.	Simple verification checks may be performed to avoid conflicting feedback. This may be because the users made an error in one of the annotation steps. These prompts could reduce the noise in the training data and help improve better models. Ideas similar to NLPReViz could be adopted for resolving potential inconsistencies in user feedback (See Figure 7(b) in Chapter 4).
Retrain	R5: Additions and deletions to the list of incidentals should be displayed to help users understand changes between model revisions.	A possible design showing a diff of the signout note revisions as the model evolves. This is similar to the output shown by *'diff'*-viewing tools showing additions, deletions, and modifications between revisions of text files. 

```
Mechanism:

35-40yof presents as level 1 trauma form scene intubated. Unknown history. Hypo

PMHx: unkn

PSHx:unkn

Meds:

All:unkn

Soc:
To Do:

NSGY cx SAH/SDH, spine f/u recs
facial trauma cx
Operative:
6/10: exlap, pericardia window, bilateral chest tubes
Injuries and Problems:
blunt chest trauma with multpile b/l rib frxs, b/l ptx s/p ct
SAH/SDH
Right temporal bone frx, clavarial fracture
Multiple thoracic and lumbar frxs and distraction injuries
c7 sponious process frx
RADS/Intervention:

CXR: left chest tube in place, multiple right rib frxs

CTH:SAH/SDH, right temporal bone frx

CT Cspine:C7 spineous process frx. right inferior occiput
```

```
Mechanism:

37 yo F who presented as a Level 1 s/p ejected MVA. Unknown history. Intubated

PMHx: unkn

PSHx:unkn

Meds:

All:unkn

Soc:
To Do:
Eventually shunt to be placed intraperitoneally
NSGY - LP 24 opening pressure, previously 27. Awaiting plans on VP shunt
ET to d/c Vac and use aquacel
Operative:
6/10: exlap, pericardial window, bilateral chest tubes (JCP/PV)
6/13: Fascial closure, gastropexy (JPI/PV)
6/14: Thoracic fusion (NSGY)
6/15: Left humerus ORIF, wrist ex-fix (Ortho)
6/16: Trach (Trauma), L distal radius ORIF (Ortho)
6/27:  Lumbar Puncture
7/3:  White Sponge wound vac
7/5: VP shunt
7/11:  Shunt exterilzation
7/14: NSGY EVD placed, thoracic right VATS/decortication
Injuries and Problems:

SAH/SDH
C7 spinous process fx
Multiple thoracic and lumbar fxs and distraction injuries
Multiple B/L rib fxs
B/L PTX
R temporal bone fx, calvarial fxs
R scapular fx
L proximal humerus, distal radius fxs
PNA/C diff
Occipital decubitus ulcer
RADS/Intervention:

CXR: left chest tube in place, multiple right rib frxs

CTH:SAH/SDH, right temporal bone frx

CT Cspine:C7 spineous process frx. right inferior occiput
```

		A view showing a log of feedback items could help keep track of user actions with options to undo them could be useful in managing model revisions.
		A progress bar showing the status of the number of documents manually annotated could help users interpret overall progress. Another useful addition could be a visualization of model performance metrics over time against a held-out set. This set could be taken from a gold-standard set or from a set of examples previously reviewed and labeled manually.

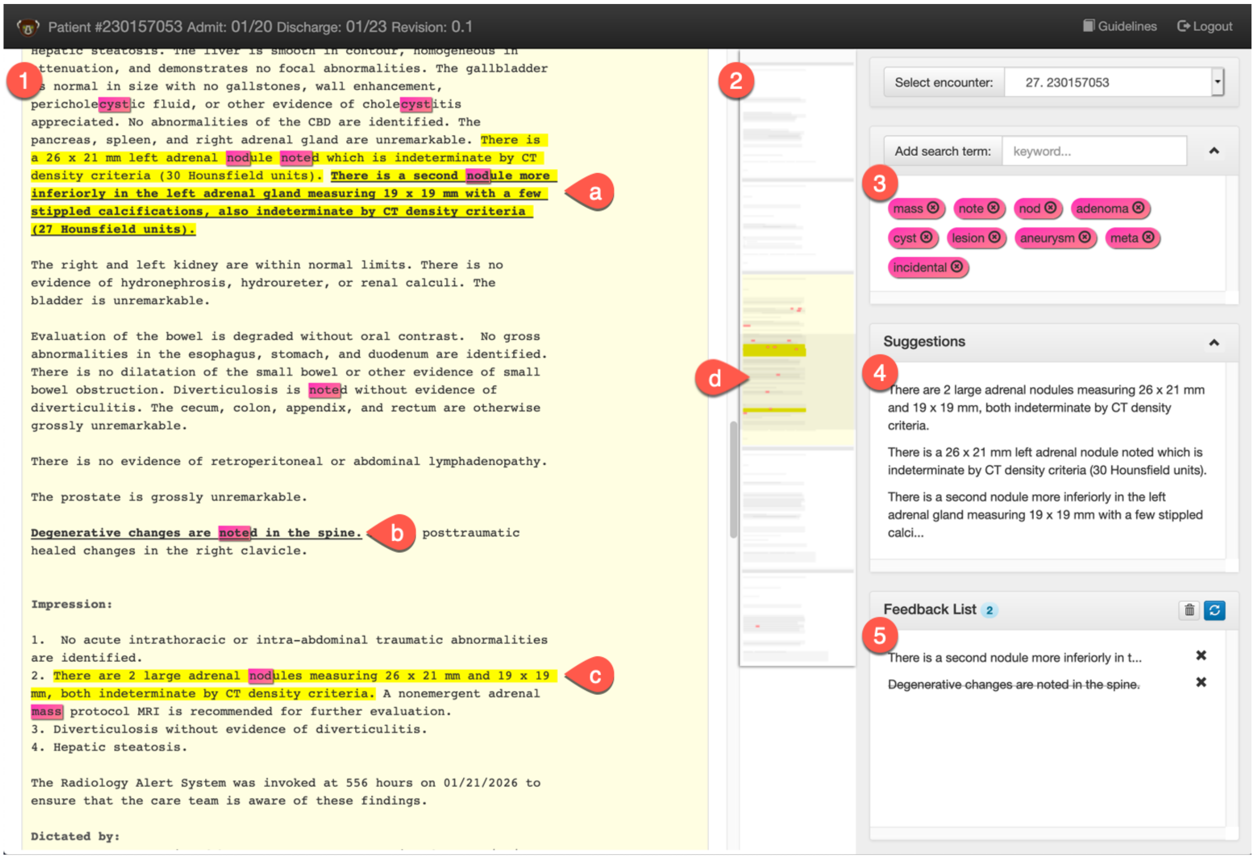
Figures 15 and 16 show the user-interface of our prototype. A video demo can be found at `http://vimeo.com/trivedigaurav/incidentals`. The following sections describe the components of the interactive feedback loop in more detail.

5.3.3.1 Review The tool presents all the radiology reports from a single patient encounter, in a continuous scrolling view. A timeline view on the top indicates the number of reports associated with the encounter and provides shortcuts to individual reports. Reports are broken into individual sections and sentences, which are marked by yellow highlights when predicted to contain incidental findings (Figure 15 1). The mini-view on the right displays an overview of the full encounter (Figure 15 2) and helps the user navigate quickly between the reports by serving as an alternate scroll bar. Varying saturation levels to draw attention to predicted incidental findings: reports with predicted incidental findings are lightly colored in yellow, followed by a darker background for sections which contains the highlighted sentence. Incidental findings are also listed in the suggestions box on the right along with a short excerpt (Figure 15 4). The user can click on these excerpts to scroll to the appropriate position in the full text report.

A list of terms relevant for identifying incidental findings, including terms such as *nodule*, *aneurysm*, *incidental*, etc. is shown in Figure 15 3. These terms are highlighted in pink in the main document and in the mini-view. Users have an option to add or remove their own terms.

5.3.3.2 Feedback To revise models, users right-click on selected text spans to launch a feedback menu enabling addition, removal, or confirmation of predicted incidental findings (Figure 16(a)). The system automatically segments the reports into sections and sentences. These can be inspected by taking the mouse cursor over them. Individual sections or sentences can be selected through a single right-click (no span selection required, Figure 16(b)). The user also has an option to specify incidental findings at the sentence, section, report, or encounter levels individually. A checked box indicates the presence of an incidental finding. Hierarchical rules are automatically applied as the user provides feedback: if the sentence is marked as an incidental then all the upper levels are also checked. A similar user action is needed to remove incorrectly predicted findings as well. The appropriate interpretation of a feedback action is inferred from the context. For example, if the only predicted sentence is removed from a section, then both the sentence as well as the section containing it are un-highlighted. Text items against which feedback is provided are bolded and underlined (Figure 15 (a) & (b)). If a user reads through a report and makes no change to predicted incidentals findings (Figure 15 (c)), the initial labels are assumed to be correct and added as implicit feedback.

5.3.3.3 Retrain A list of all current feedback is provided on the bottom panel of the right sidebar (Figure 15 5), which shows a short excerpt from each selected text span. If a user removes highlighted incidental findings, these are also listed in the sidebar and are denoted by a strike through. Clicking on these items in the feedback list scrolls the full-text note to appropriate location. The 'x'-button allows the users to undo feedback actions and remove them from the feedback list. Switching to different patient encounter triggers model retraining. Once the retraining is complete, the new predictions are highlighted. The refresh button can also be used to manually re-train and refresh predictions.



Patient #230157053 Admit: 01/20 Discharge: 01/23 Revision: 0.1

Hepatic steatosis. The liver is smooth in contour, homogeneous in ttenuation, and demonstrates no focal abnormalities. The gallbladder s normal in size with no gallstones, wall enhancement, pericholecystic fluid, or other evidence of cholecystitis appreciated. No abnormalities of the CBD are identified. The pancreas, spleen, and right adrenal gland are unremarkable. There is a 26 x 21 mm left adrenal nodule noted which is indeterminate by CT density criteria (30 Hounsfield units). **There is a second nodule more inferiorly in the left adrenal gland measuring 19 x 19 mm with a few stippled calcifications, also indeterminate by CT density criteria (27 Hounsfield units).**

The right and left kidney are within normal limits. There is no evidence of hydronephrosis, hydroureter, or renal calculi. The bladder is unremarkable.

Evaluation of the bowel is degraded without oral contrast. No gross abnormalities in the esophagus, stomach, and duodenum are identified. There is no dilatation of the small bowel or other evidence of small bowel obstruction. Diverticulosis is noted without evidence of diverticulitis. The cecum, colon, appendix, and rectum are otherwise grossly unremarkable.

There is no evidence of retroperitoneal or abdominal lymphadenopathy.

The prostate is grossly unremarkable.

Degenerative changes are noted in the spine. posttraumatic healed changes in the right clavicle.

Impression:

1. No acute intrathoracic or intra-abdominal traumatic abnormalities are identified.
2. There are 2 large adrenal nodules measuring 26 x 21 mm and 19 x 19 mm, both indeterminate by CT density criteria. A nonemergent adrenal mass protocol MRI is recommended for further evaluation.
3. Diverticulosis without evidence of diverticulitis.
4. Hepatic steatosis.

The Radiology Alert System was invoked at 556 hours on 01/21/2026 to ensure that the care team is aware of these findings.

Dictated by:

Select encounter: 27. 230157053

Suggestions

There are 2 large adrenal nodules measuring 26 x 21 mm and 19 x 19 mm, both indeterminate by CT density criteria.

There is a 26 x 21 mm left adrenal nodule noted which is indeterminate by CT density criteria (30 Hounsfield units).

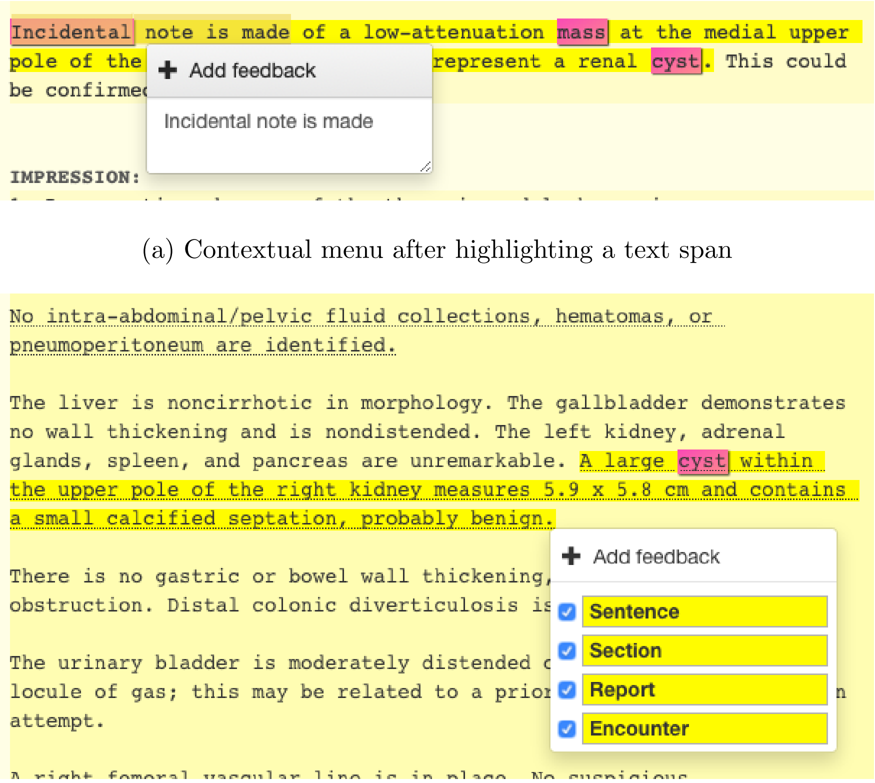
There is a second nodule more inferiorly in the left adrenal gland measuring 19 x 19 mm with a few stippled calci...

Feedback List 2

There is a second nodule more inferiorly in t... ✖

~~Degenerative changes are noted in the spine.~~ ✖

Figure 15: **1) A de-identified radiology report of CT imaging in a patient with trauma. It revealed a *nodule* as an incidental finding that is highlighted in yellow by the prototype tool (a & c). The users are able to add incidentals missed by the prototype (bolded in a) and also remove incorrectly highlighted findings (b). 2) The tool shows an overview of the patient case in a miniaturized view of all the records with highlights marking regions of interest (d). In the right sidebar, the tool allows the users to define search terms to be highlighted in pink 3) These can be seen as rules which can help attract user attention to potentially important parts of the case. 4) Shows the list of predictions made by the system. Clicking on a blurb item scrolls the report view to relevant prediction into view. 5) Shows a log of feedback items and changes recorded by the user.**



Incidental note is made of a low-attenuation mass at the medial upper pole of the

represent a renal cyst. This could be confirmed

+ Add feedback

Incidental note is made

IMPRESSION:

(a) Contextual menu after highlighting a text span

No intra-abdominal/pelvic fluid collections, hematomas, or pneumoperitoneum are identified.

The liver is noncirrhotic in morphology. The gallbladder demonstrates no wall thickening and is nondistended. The left kidney, adrenal glands, spleen, and pancreas are unremarkable. A large cyst within the upper pole of the right kidney measures 5.9 x 5.8 cm and contains a small calcified septation, probably benign.

There is no gastric or bowel wall thickening, obstruction. Distal colonic diverticulosis is

The urinary bladder is moderately distended c locule of gas; this may be related to a prior n attempt.

+ Add feedback

☑ Sentence

☑ Section

☑ Report

☑ Encounter

(b) Shortcut menu to provide feedback on the entire sentence

Figure 16: **(a) Users can add feedback by highlighting a span of text and triggering the contextual menu with a right click. (b) To add or remove an entire sentence, report or encounter (patient-case), the contextual menu can also be launched without manually selecting any text spans.**

5.3.4 Implementation and Deployment

The prototype system is implemented as a client-server architecture. The user interface is built using AngularJS (angularjs.org) framework. The learning pipeline is implemented as Falcon (a web API framework for Python; falconframework.org) web app in python. Pre-processing steps such as sentence segmentation were performed using spaCy (a Python NLP library; spacy.io [92]). We used MongoDB (a NoSQL database; mongodb.com) to store pre-processed text along full-text reports. This architecture allowed us to perform quick re-training on the fly without any delays that were noticeable to the users. Support vector machine models were built using *scikit-learn* (a machine learning library for Python; scikit-learn.org [94]). A list of other packages and dependencies is available along with our tool's source code at github.com/trivedigaurav/lemr-vis-web and github.com/trivedigaurav/lemr-nlp-server.

5.3.5 Evaluation

Interactive Machine Learning systems require an evaluation from two different perspectives [95, 42]: model performance and system usability. Thus, we divided the evaluation into two parts by mapping them to the two sub-hypotheses discussed in Section 5.2 (H1: asking questions about usability and H2: measuring the efficiency and correctness of the models). We recruited physicians with experience in reading radiology notes and identifying incidental findings, to participate in our study. Using Friedman's study type definitions [96] this evaluation falls under the 'Lab Study' setting. Our study protocol was approved by the University of Pittsburgh's Institutional Review Board (PRO18070517). The results from this study are discussed in Section 5.4.

We recruited 15 participants with a degree in medicine and training in critical care, internal medicine, or radiology. They were given a $50 gift card as compensation for participating in the study over web-conferencing. Before each study session, we collected general background information about the participants, their clinical experience and their knowledge of using NLP tools. We went over the annotation guidelines and allowed the participants to seek any clarifications. The participants were free to ask questions about the guidelines throughout the study. After a short walkthrough of the prototype, they did a trial run of the tool before they reviewed actual cases.

The participants were then asked to review radiology reports from the de-identified dataset described above and identify incidental findings. We interleaved the encounter presentation in control and intervention conditions. In the intervention condition, we enabled all predictive features of the prototype about the incidentals. However, for the control case, while the user feedback was saved to revise models, the users were not shown any predictions to simulate the existing practice for documenting incidentals. The participants to free to review as many patients they could in the stipulated time span. We logged time spent on each patient case along with their interactions with the tool.

After 60 minutes of reviewing notes, we presented the participants with a post-study questionnaire. It included questions about the issues faced by them and prompts to encourage their feedback on individual design components of prototypes.

5.3.5.1 Usability evaluation The goal of the usability evaluation is to demonstrate overall usability and usefulness of the tool. We performed a System Usability Scale (SUS) [69] based evaluation along with think-aloud sessions and semi-structured interviews. SUS offers a quick and reliable measure for overall usability. It asks 10 questions with 5-point Likert Scale responses ranging from 'Strongly Agree' to 'Strongly Disagree'. The responses to these questions are used to compute a SUS score between 0 to 100. We also recorded subjective feedback about individual components of the prototype to gather feedback about subsequent versions of such a tool.

5.3.5.2 Evaluating correctness We evaluated efficiency and model accuracy through a combination of intrinsic and extrinsic approaches:

1. *Intrinsic Evaluation:* A comparison of how our system predicts important information with human-annotated test data (130 encounters, 694 reports). We used F1, Precision and Recall as our evaluation metrics for the intrinsic evaluation. The models were bootstrapped by training an initial model on a set of 6 patient encounters. 2/3 of the dataset was used for review during the study and 1/3 of the cases were held out for testing.

The distributions of positive incidentals were similar for test and development tests at all three levels. The same test and train split was used for all participants to allow comparison of final results.

2. *Extrinsic Evaluation:* We measured the time spent per patient case, as well as the total number of user-actions in the control and intervention conditions. We split half of the encounters in the development set into control and intervention conditions for each participant. We shuffled these lists of encounters between each run of the study. Since each participant was presented the control and intervention condition in an interleaved manner, we obtained paired samples for each participant. The first encounter from each of the conditions was ignored for timings calculations to minimize learning effects. We observed that most users were able to clarify any questions or concerns about the interface after the initial trial run and the two patient cases during the actual study.

5.4 RESULTS

5.4.1 Quantitative Analysis

Table 9 gives a summary description of our participants. We computed an average SUS score of 78.67 out of 100. A SUS score of 68 is considered an average usability performance [97].

Table 9: ***Participants*: Description of user study participants. An average SUS (System Usability Scale) score [69] of 78.67 was observed using post-study questionnaires.**

	Position	**Years in position**	**Area**	**Role**	**Experience with NLP?**	**SUS Score**

p1	Physician	<5 yrs	Pediatric Emergency Medicine	Clinician	No	72.5
p2	Resident	<5 yrs	General Surgery	Clinician, Researcher	No; Involved in a past project	80
p3	Resident	<5 yrs	Radiology	Clinician	No; But familiar	87.5
p4	Resident	<5 yrs	Radiology	Clinician	No	62.5
p5	Resident	<5 yrs	Neuro-radiology	Clinician, Researcher	No	77.5
p6	Resident	<5 yrs	Radiology	Clinician	No	85
p7	Resident	<5 yrs	Internal Medicine	Clinician	No	92.5
p8	Doctoral Fellow	<5 yrs	Biomedical Informatics	Researcher	No	95
p9	Asst. Professor	<5 yrs	Internal Medicine	Clinician	No	67.5
p10	Resident	5-10 yrs	General Surgery	Clinician	No	75
p11	Resident	5-10 yrs	Critical Care	Clinician	No	70
p12	Research Staff	<5 yrs	Biomedical Informatics	Clinician, Researcher	No	77.5
p13	Senior Research Scientist	10+ yrs	Biomedical Informatics	Researcher	No	80

p14	Asst. Professor	10+ yrs	Internal Medicine	Clinician	No	87.5
p15	Resident	<5 yrs	General Surgery	Clinician	No	70

The participants reviewed between 12 to 37 cases (mean=29.33). The changes in F1 scores on the test dataset (relative to the gold-standard labels) at each revision are shown in Figures 17-19. Comparing the F1 scores from the against the initial bootstrapped model to the final models built by participants in the hour-long session, we observed an increase from 0.31 to 0.70–0.79 (mean=0.75) for reports, 0.32 to 0.57–0.73 (mean=0.68) for sections and from 0.22 to 0.50–0.68 (mean=0.60) for sentences. Table 10 shows a Precision, Recall and F1 comparisons between initial and final models.

Agreement of feedback labels relative to the gold-standard labels ranged from Cohen's κ of 0.76–0.95 for reports, 0.84–0.96 for sections, and 0.74–0.91 for sentences.

We observed statistically significant lower time in intervention encounter as compared to the control (mean time: 134.38 vs. 148.44 seconds, Wilcoxon: $Z = 10$, $p < 0.005$). Average time per each patient case is shown in Figure 20 for each participant.

While comparing the total number of feedback actions, we observed statistically significant lower feedback counts in intervention condition (average counts: 42 vs. 55.07, Wilcoxon: $Z = 13.5$, $p < 0.05$). See Figure 21.

We found no statistical differences between final F1 scores or agreement with gold standard labels between control and intervention conditions at any level (Figure 22).

5.4.2 Qualitative Analysis

Open-ended subjective feedback was mainly positive for our tool and recorded no major usability problems: *"intuitive and easy to use after initial training"*. Overall the idea for

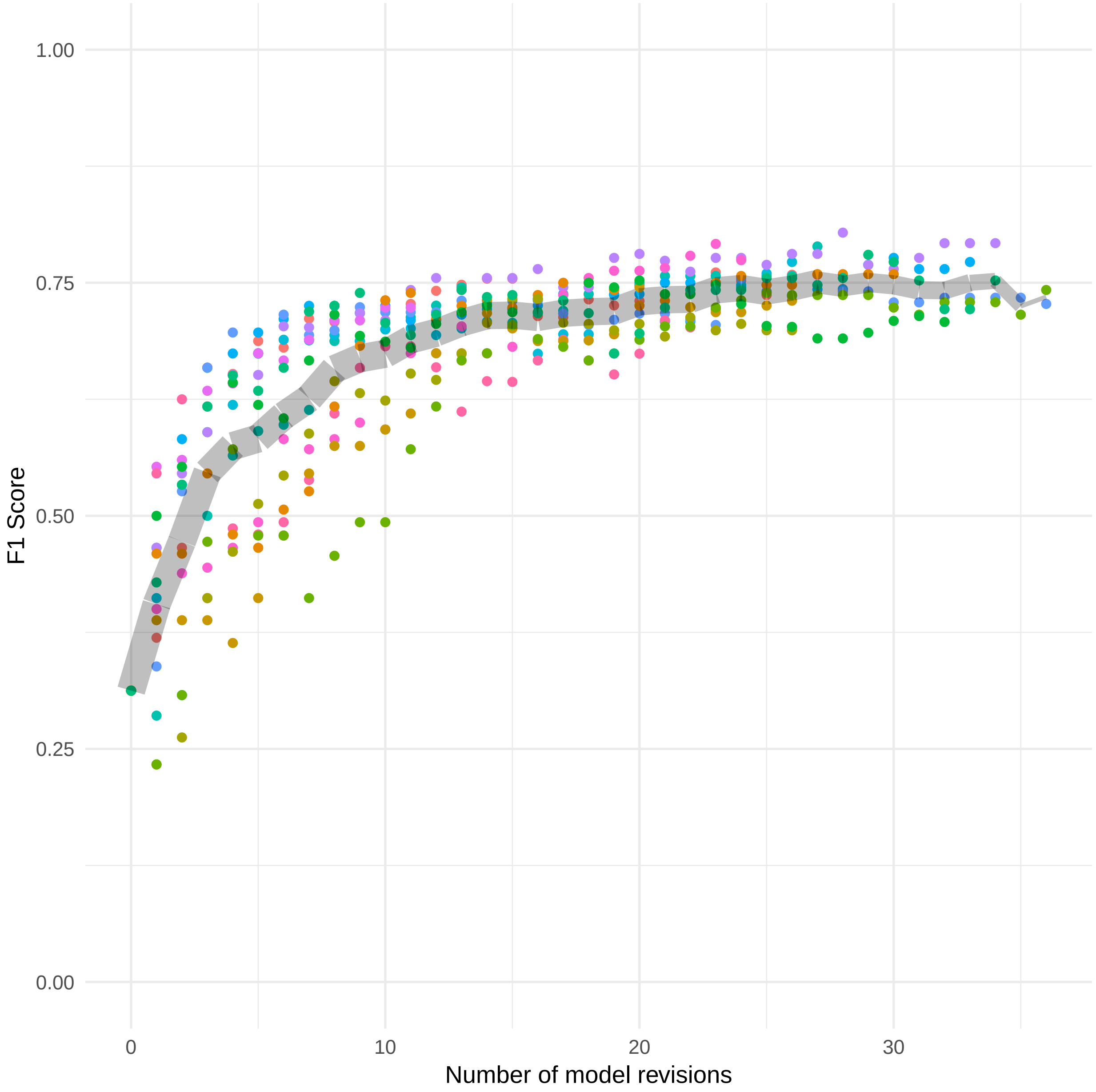


Figure 17: ***Reports:*** **Change in F1 scores over time at the report level. The colored points represent individual participants. The grey band marks the average score and tapers off in thickness to represent the number of participants completing that revision during the study.**

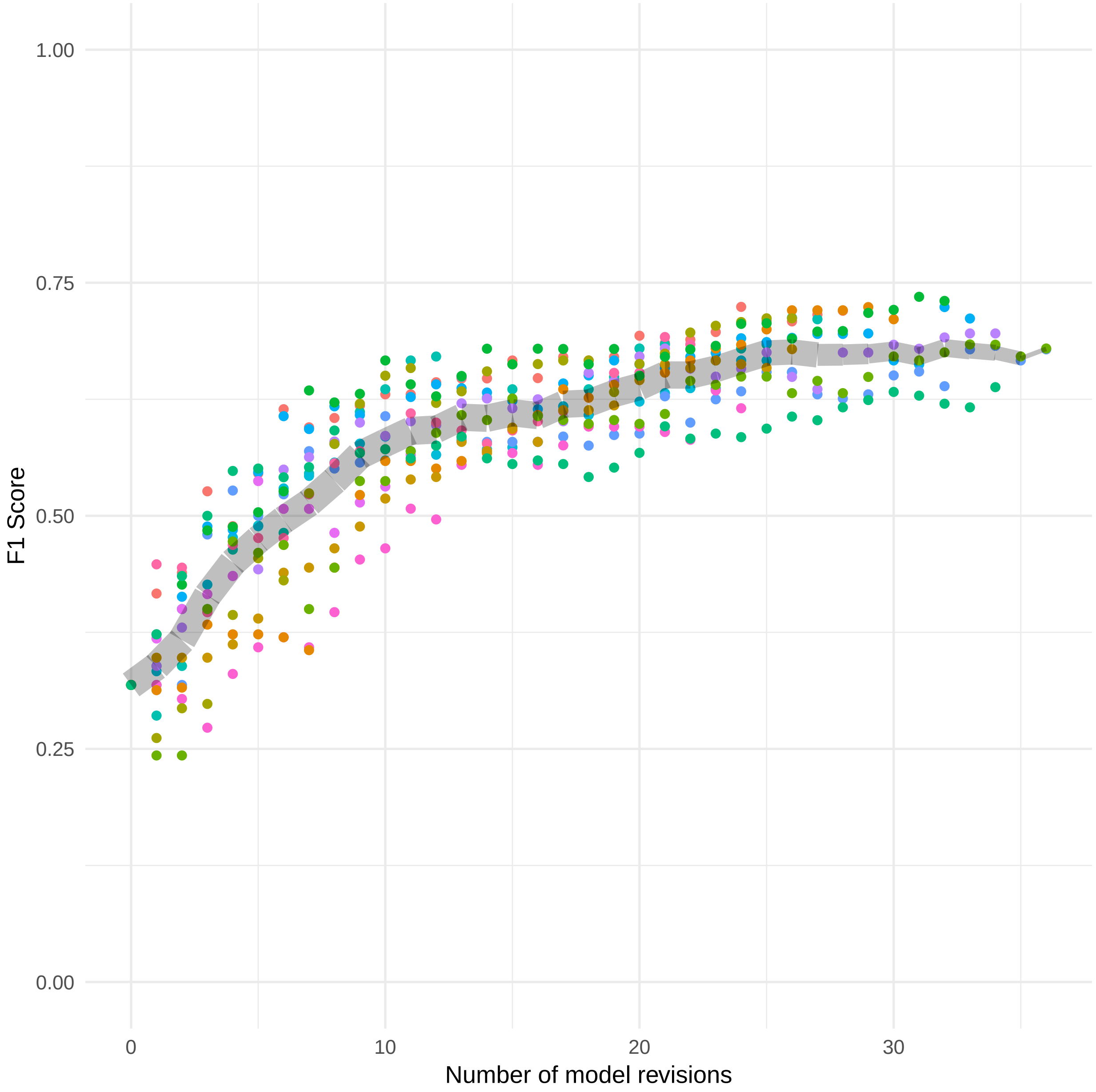


Figure 18: ***Sections:*** **Change in F1 scores over time at the section level. The colored points represent individual participants. The grey band marks the average score and tapers off in thickness to represent the number of participants completing that revision during the study.**

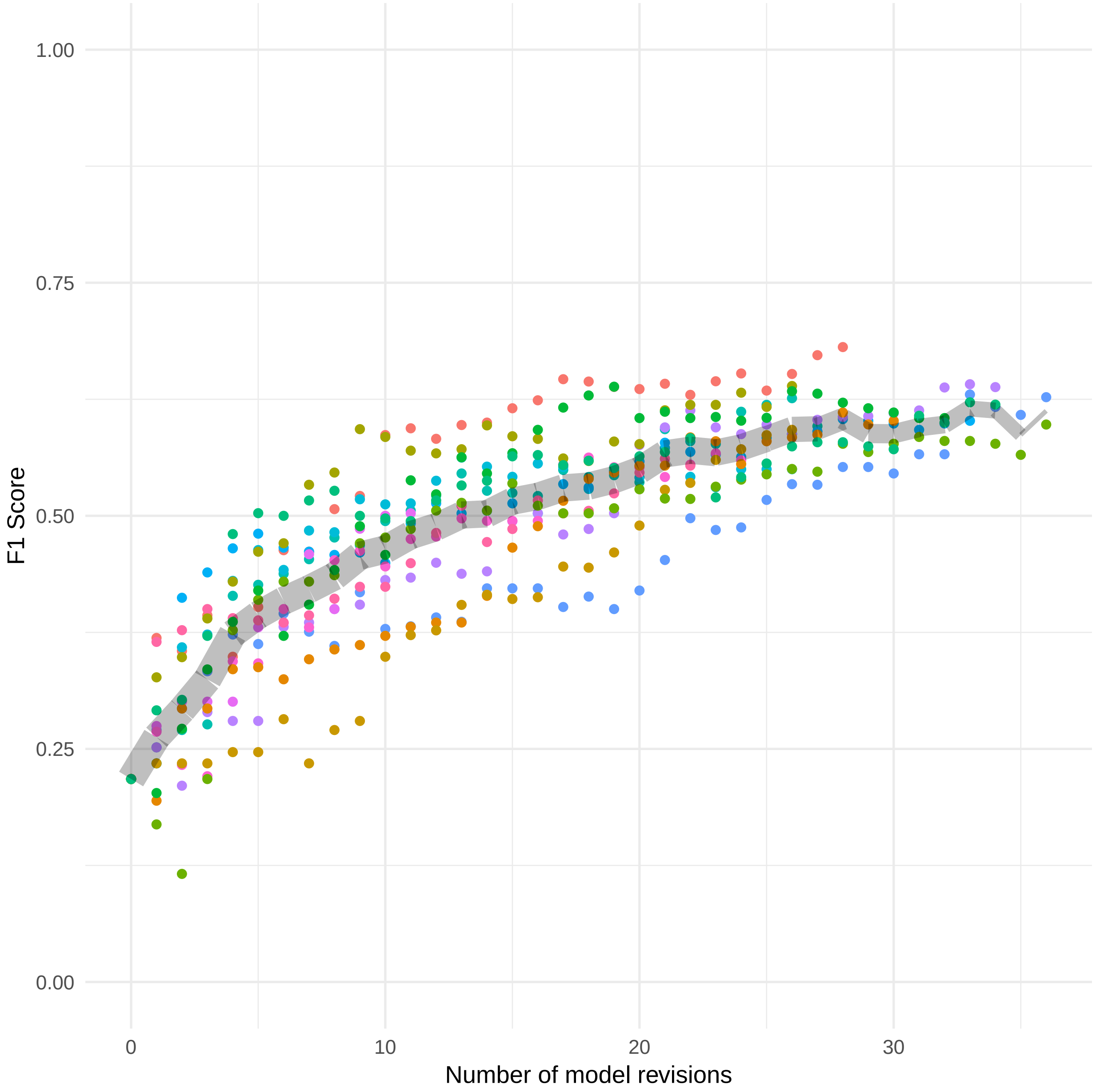


Figure 19: ***Sentences:*** **Change in F1 scores over time at the sentence level. The colored points represent individual participants. The grey band marks the average score and tapers off in thickness to represent the number of participants completing that revision during the study.**

Table 10: ***Final scores*: Comparison of Precision (P), Recall (R) and F1 scores between initial and final model revisions for all 15 participants. Cohen's Kappa (κ) measures the agreement between feedback provided by the participants and the gold standard labels. The initial model was trained on the same 6 patient cases to bootstrap the learning cycle.**

	Reports				**Sections**				**Sentences**			
	P	**R**	**F1**	κ	**P**	**R**	**F1**	κ	**P**	**R**	**F1**	κ
Initial	0.90	0.19	0.31		0.86	0.20	0.32		0.84	0.13	0.22	
p1	0.70	0.79	0.74	0.76	0.76	0.68	0.72	0.83	0.76	0.62	0.68	0.74
p2	0.77	0.81	0.79	0.81	0.80	0.64	0.71	0.85	0.78	0.48	0.60	0.75
p3	0.81	0.74	0.77	0.84	0.80	0.55	0.65	0.87	0.76	0.43	0.55	0.80
p4	0.81	0.74	0.77	0.95	0.82	0.63	0.71	0.94	0.79	0.48	0.60	0.87
p5	0.70	0.75	0.73	0.89	0.75	0.62	0.68	0.90	0.75	0.54	0.63	0.79
p6	0.79	0.79	0.79	0.92	0.81	0.61	0.70	0.96	0.82	0.52	0.64	0.88
p7	0.83	0.64	0.72	0.93	0.77	0.45	0.57	0.87	0.84	0.36	0.50	0.84
p8	0.90	0.68	0.77	0.92	0.86	0.48	0.62	0.91	0.83	0.42	0.56	0.85
p9	0.80	0.62	0.70	0.89	0.84	0.58	0.68	0.94	0.81	0.42	0.55	0.86
p10	0.75	0.77	0.76	0.91	0.80	0.64	0.71	0.91	0.79	0.48	0.60	0.82
p11	0.75	0.72	0.73	0.91	0.77	0.61	0.68	0.93	0.78	0.48	0.59	0.83
p12	0.72	0.68	0.70	0.90	0.74	0.68	0.71	0.90	0.77	0.55	0.64	0.75
p13	0.82	0.68	0.74	0.87	0.83	0.58	0.68	0.85	0.88	0.45	0.60	0.76
p14	0.67	0.75	0.71	0.84	0.81	0.66	0.73	0.90	0.81	0.48	0.60	0.81
p15	0.79	0.72	0.75	0.92	0.73	0.57	0.64	0.95	0.88	0.48	0.62	0.91
Mean	0.77	0.73	0.75	0.88	0.79	0.60	0.68	0.90	0.80	0.48	0.60	0.82
SD	±.06	±.06	±.03	±.05	±.04	±.07	±.04	±.04	±0.04	±.06	±.04	±.05

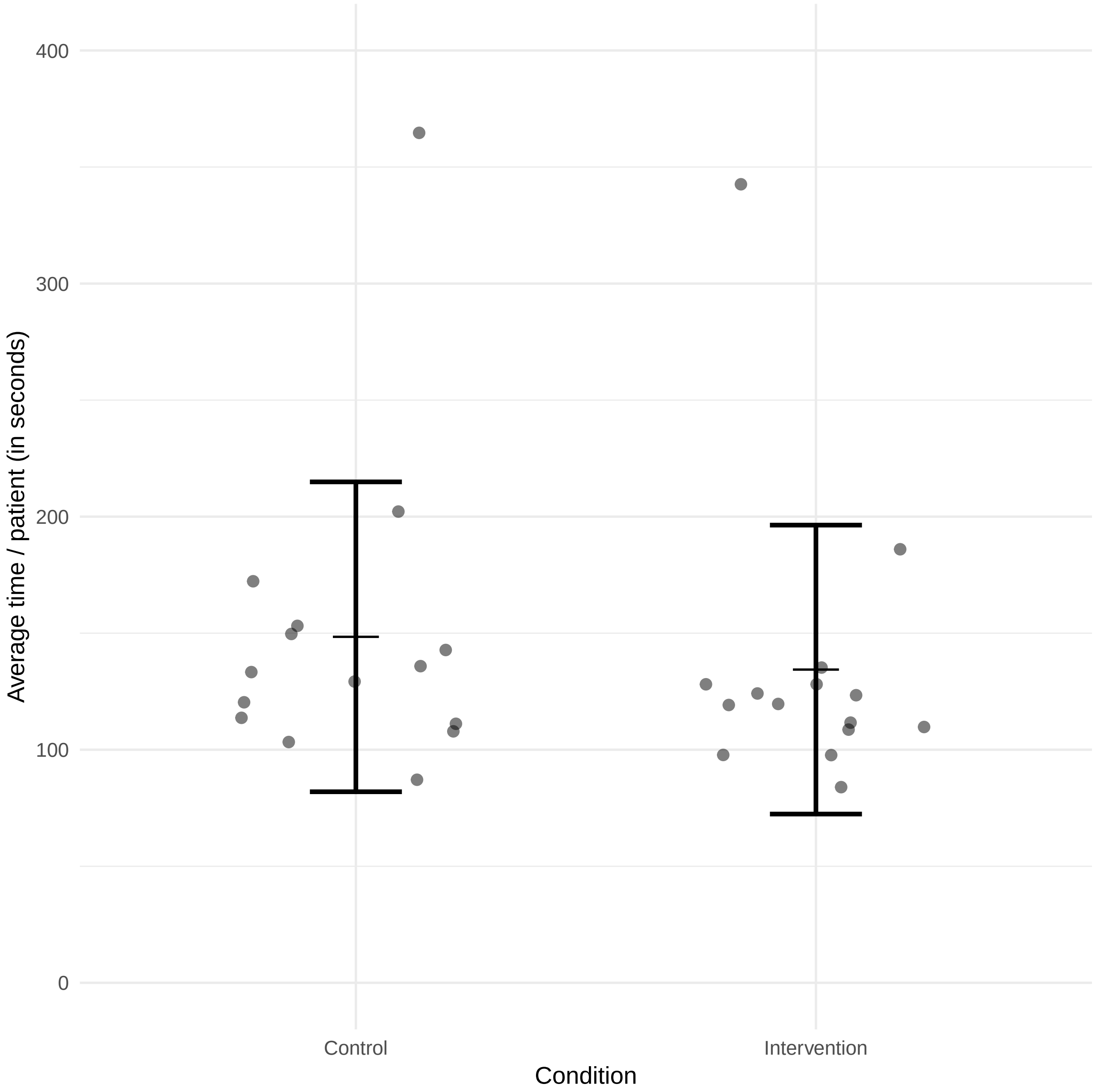


Figure 20: **Average time spent in seconds between control and intervention conditions. The dots represent individual samples. We observed a statistically significant lower time in intervention vs control conditions (mean time: 134.38 vs. 148.44 seconds, Wilcoxon: $Z = 10$, $p < 0.005$). One participant spent much longer time per patient case than others and can be seen as an outlier in both the conditions.**

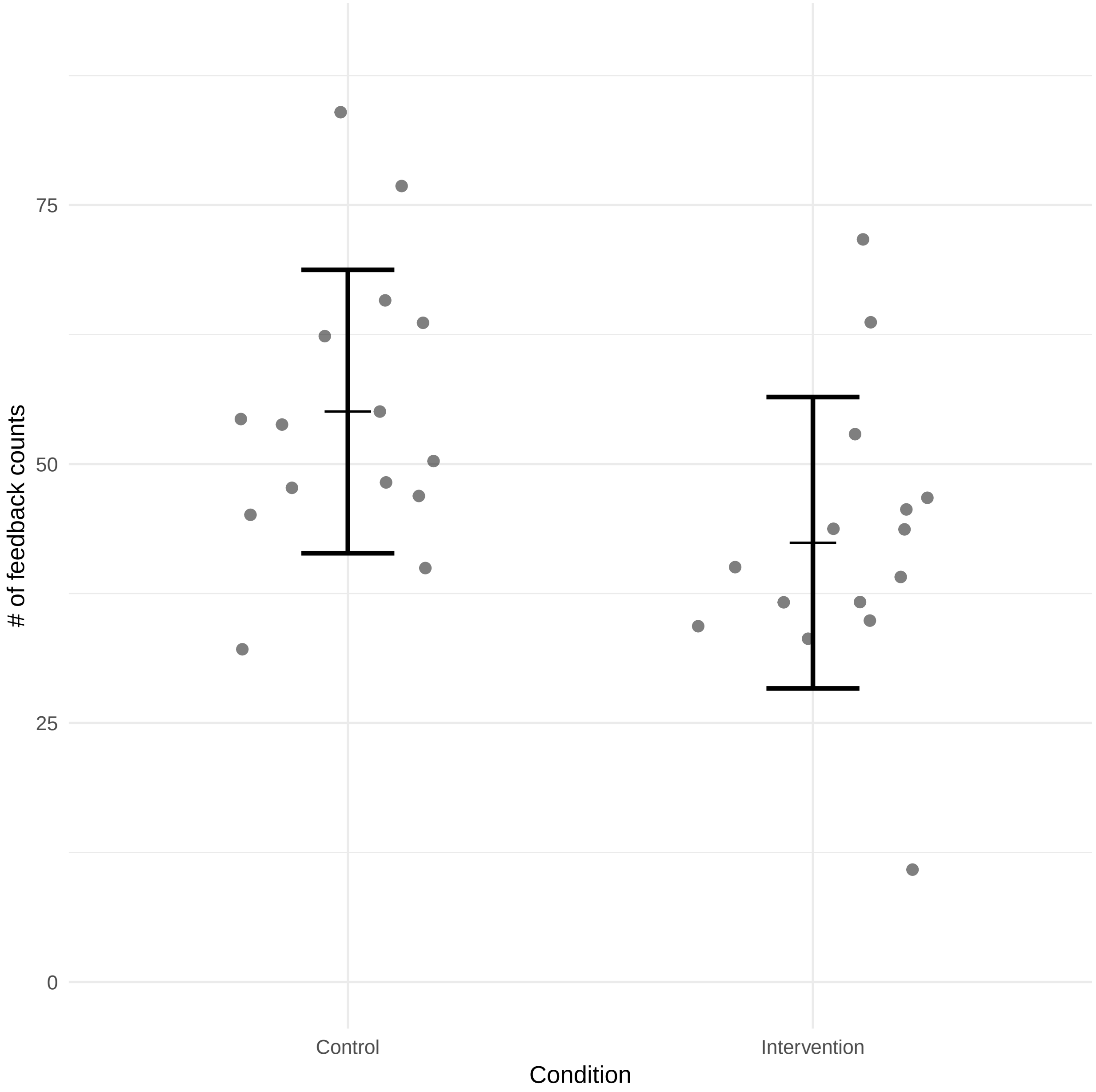
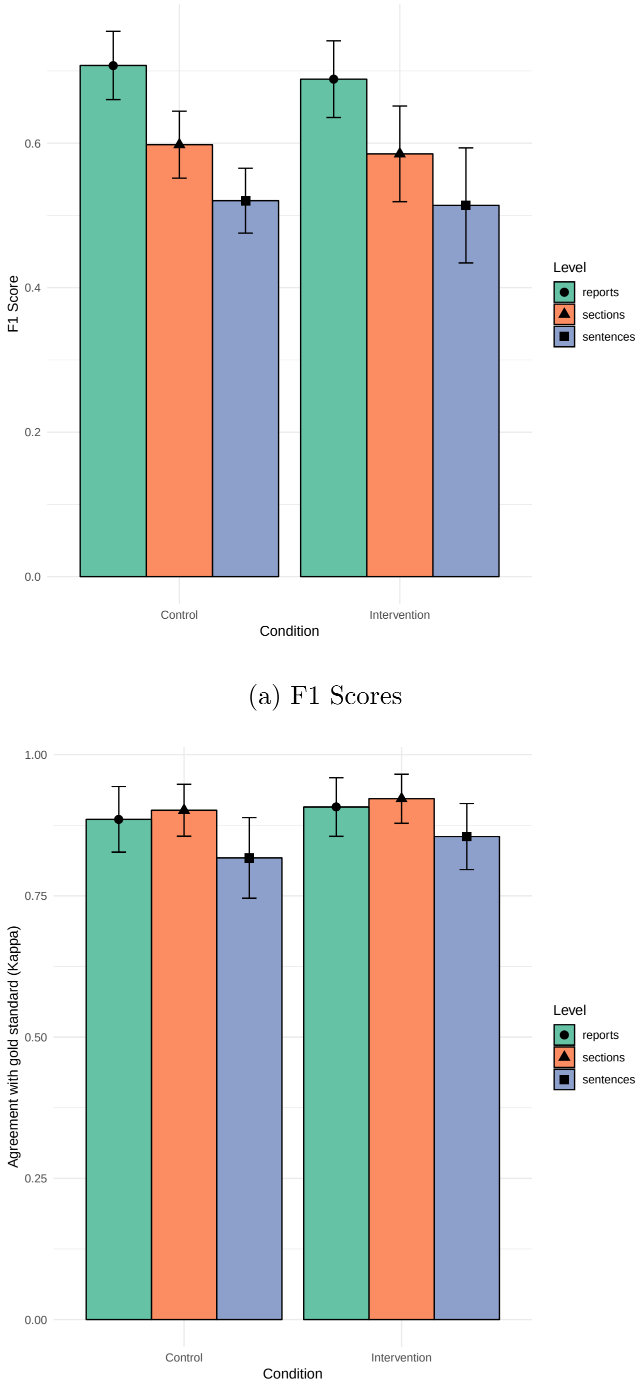


Figure 21: **Average feedback counts between control and intervention conditions. The dots represent individual samples. We observed statistically significant counts in the intervention vs. control conditions (average counts: 42 vs. 55.07, Wilcoxon:** $Z = 13.5$, $p < 0.05$)



(a) F1 Scores

(b) Kappa scores

Figure 22: **We found no statistical differences between final F1 scores or agreement with gold standard labels between control and intervention conditions at any level.**

highlighting incidental findings was well received:

> *"In my personal practice, I have missed out on incidental findings* [on occasion] *... if we are able to highlight them, it would be very helpful."*

> *"It's useful to verify that I didn't miss anything."*

5.4.2.1 Review Participants appreciated the report view which provided easy access to all the related scans, *"In the system that I use* [at work], *you have to open each report individually rather than having to see them at once and scroll through them easily."*

All participants found it useful to be able to define search terms that were highlighted in pink (Figure 15 3). While we provided the functionality to add and remove custom terms most participants did not make use of that feature.

Participants praised the highlighting components of the tool as well, *"I liked the highlighting a lot... When it was already highlighted, my response to confirming that was an incidental was faster"*. Highlighting on reports, sections and sentences in increasing saturation levels was also found to be useful: *"I knew I was heading towards a highlighted page... made me focus more."*, [it signaled] *"...that there is something going on."*

Most users didn't pay attention to the mini-patient view but acknowledged that it would be useful in real-world use cases. But a small group of users used it extensively: *"Made it easy to see where incidentals have been found"*, *"Helped me understand which page of the record I am at"*

5.4.2.2 Feedback Users found the mechanism for providing feedback straightforward (Figure 16). Right click and highlight (Figure 16(a)) was useful when sentence boundary detection had issues: *"There were some examples when I did want the whole sentence to be highlighted."*

All but one participant gave feedback only at the sentence level even though the tool allowed them to provide feedback at the report and section levels as well.

User perception of the feedback list on the bottom right was mixed (Figure 15 5). While some participants made extensive use to undo feedback actions, others didn't pay enough attention since it did not occupy a prominent location on the screen. One participant suggested that this could be combined in a single box along with system suggested incidentals (Figure 15 3), while another insisted that they occupy separate views: *"This was helpful because sometimes I noticed that I highlighted too much, so I could go back and fix it."*

5.4.2.3 Retrain Most participants agreed that while shortcuts to click on incidental blurbs and jump to those findings in the text would be useful in a real-world scenario, they did not use that feature during the study. Several participants remarked that they didn't get a chance to explore every component of the user interface as they were mainly focused on the study task of reviewing notes.

> *"I picked up more speed towards the end"*
>
> *"If I regularly used this tool then it may be even more useful in skimming through the text – saves a lot of time"*

5.4.2.4 Future Directions Table 11 summarizes the list of design improvements suggested by the participants. Participants also had suggestions for the tool beyond the incidentals use case.

> *"We scan through a lot of reports and notes, so it would be very to help to identify important findings from the rest of the noise, ...* [such a tool] *could potentially help us streamline a lot of our workflow."*

Depending on the situation, the clinicians are looking for specific types of problems:

> *"If I see bruising... I may go back and see what the radiologist noted about injuries."*

Besides incidentals, interactive NLP could be used to build models for other kinds of findings such as injuries, effusions and clinically relevant things that may have an impact on a patient's care and treatment. Participants also pointed out use-cases in radiology. For example, such as a system may be used to remind radiologist about missed incidentals when they dictate a report. Based on the findings listed in the report, the system could auto-suggest relevant findings to mentioned in impression including a recommendation for a follow-up based on the current guidelines. Other suggestions stemmed from use-cases in

Table 11: **List of design recommendations for improving the system from the user study.**

Category	**Recommendation**
Review	1. Allow users to define their custom color schemes for highlights.
	2. Include negation rules for keyword search. For example, differentiate between: *'mass'* and 'no *mass*' [98].
	3. Enable top feature highlighting as explanations for the predictions.
	4. Distinguish between different kinds of sections in the reports (eg. *Impression* and *Findings* vs. other sections). Allows users to quickly jump to specific sections.
Feedback	1. All but one participant gave feedback only at the sentence level even though the tool allowed them to provide feedback at report and section levels as well. Feedbacks may be provided with a single right click instead of triggering a contextual menu first. Options for other levels may then be provided with a pop-up menu over these highlighted feedback items.
	2. Display intelligent blurbs in the feedback list that drew attention to the main findings or keywords (e.g. *'mass'* or *'nodule'*) instead of just the leading part of the sentence.
Re-Train	1. Allow some free-form comments along with the feedback marking incidentals. Not only this can serve as a helpful annotation for the other members of the team, the learning pipeline may also use that as an additional input to improve models.
	2. Some of the pre-defined search keywords (in pink) raised a lot of false-positives (eg. *'note'*). An automated mechanism to suggest addition and removal of these terms may be useful.

reading pathology reports, blood reports, labs, etc. Another participant while acknowledging the benefits of AI to support clinical workflow also added a caveat about potential bias due to automation:

> *"Clinical notes have a lot of text and are hard to read and having something that highlights a finding – everything that saves time is helping me do the job better. Although I wouldn't want to miss something if it is not highlighted by the too."*

5.5 DISCUSSION

In this work, we demonstrated how our interactive learning framework defined in Chapter 4 can be used in a clinical care setting. We developed a prototype tool to help clinicians identify incidental findings from trauma scans. This can be seen as a step beyond the document classification task in NLPReViz. Here, we are also interested in identifying relevant portions within a document as well. Our prototype combines interactive displays of NLP results with tools for reviewing text and revising models.

Our user study demonstrated successful use of the prototype by our intended users as they built NLP models bootstrapped from a small number of initial examples. Our users were clinicians with little or no experience with building machine learning models. We observed an average increase in F1 score from 0.31 to 0.75 for reports, 0.32 to 0.68 for sections, and from 0.22 to 0.60 for sentences (Table 10) over the 60 minutes long study. Specifically, we observed large improvements in our recall scores between the initial and final models. We recorded an average increase of 0.19 to .72±.05 for reports, .20 to .60±.07 for sections, and .13 to .48±.06 for sentences. Precision and recall scores were balanced for reports, but sections and sentence had lower recall scores. This may be due to heavily skewed training data (Table 7). From our extrinsic evaluation, we found that tool helped significantly reduce the time spent for reviewing patient cases (134.30 vs. 148.44 seconds in intervention and control respectively) while maintaining the overall quality of labels measured against our gold-standard. This was because the participants needed less time identifying and marking incidentals in the intervention condition where the tool had already highlighted them. We also measured a very good usability performance with an overall SUS score of 78.67. This is considerably above an average usability score of 68.

Subjective feedback about our user interface was also very positive. We compiled a list of participant feedback from the study for future design revisions. The users suggested some extensions to our work and how such a tool may be applied to support other clinical workflows in Table 11. We also observed that the restricted duration of the study and focus on the study tasks prevented participants to attend to all features of the prototype exhaustively. Participants comments that for features such as 'search-term highlights' and the 'mini-patient view,' they found little time to explore. Future exploration may involve designing study tasks that tease out how individual components affect usability. Users relied almost exclusively on feedback given at the sentence level. This is not surprising, as most incidental findings are succinctly described in a single sentence. We expected that the main application of section and report level highlighting would be for the identification of false positives. While the tool automatically highlighted segmented sentences for providing feedback, participants also sometimes manually highlighted parts of sentences instead when sentence boundary detection faltered. Deeper investigation into usage patterns and resulting models might provide some insight into which factors influenced user actions, and how they might be resolved in future redesigns. For the retrain step, we did not address the problem of measuring model performance over time. Future work may explore visualization of performance metrics against a held-out set and examples already reviewed and labeled manually by the users.

Physicians spend a large proportion of their time searching notes and reports to learn relevant information about patients. Although our work focused on the use of incidental findings as an example use case, the problem of identifying important or relevant information from free-text reports may be generalized for many similar applications including preparing discharge summaries, formulating reports for rounding, and authoring consultation notes. Several of these applications were suggested by the study participants. Once of the direct extension of this current prototype would involve predicting sentences for all sections of the complete signout note. This would involve keeping track of multiple model-types for each section in the note. The users would then be highlight and mark sentences to be referenced in these different sections. A mock-up of such an interface is shown in Table 8 *(Feedback)*.

One of the limitations of our study is the lack of access to the real EMR systems for comparison. Although the clinicians appreciated our cleaner design, the interface was designed solely for the user-study task and not as a general purpose EMR. We simulated the existing workflow in traditional EMRs by hiding NLP predictions in the control condition. We also found that the user study defined a slightly artificial task as the clinicians reviewed many patients at once. In a real-world scenario, clinicians may review notes for many different objectives together at once and not for a singular task such as identifying incidentals.

Although many users supported the need for such a tool from their prior experience with missing out on incidentals while reading scans.

We also used simpler machine learning pipelines as a trade-off for faster speed and easier implementation vs. classification performance, in order to demonstrate our interactive approach. Future work may involve an exploration of more recent modeling approaches for classifying incidentals (such as in [93]). For example, we may design mechanisms for handling noisy labels, and consider soft-labels based on user expertise. Other unaddressed issues include easier management of user models, including allowing version rollbacks and keeping track of model performance. A progress bar showing the status of the number of documents manually annotated could help users interpret overall progress. Another useful addition could be a visualization of model performance metrics over time using a held-out set and/or the set of training examples. The examples could be taken from a gold-standard set or from the set of examples previously reviewed and labeled manually.

Our user study supports the viability of adopting interactive NLP tools in clinical care settings. We used incidentals as an example use-case in our work, but the problem of identifying important or relevant information from free-text reports can be generalized for many practical applications. The incidentals problem also demonstrates the need for highly customized NLP models depending on different clinical settings: specialty, team, accepted guidelines, etc. and objectives. The guidelines for labels may also evolve over time which makes it challenging to maintain models centrally. This further bolsters the argument in favor of introducing interactive learning that allows the consumers of the NLP models to review and revise models. We are also addressing the problem of lack of upfront labeled training data by building tools that integrate machine learning into a clinician's workflow. By building interactive NLP tools that focus on the clinicians as end-users, we may be able to realize the true potential of using NLP for real-world clinical applications.

6.0 CONCLUSION

Advances in machine learning have resulted in a renewed interest in the use of artificial intelligence in medicine and healthcare. This can be seen by the recent surge in the number of peer-reviewed studies as well as the rising number of FDA approvals for AI applications in medicine [99]. This has the potential for remarkable improvements in clinicians' workflow and productivity, and also patient outcomes. A majority of these applications deal with clinical image interpretation tasks in a variety of biomedical application in radiology, pathology, dermatology, etc. While free-text notes contain rich information about a patient, real-world applications of NLP on them remain few and far between. Despite advances in Natural Language Processing (NLP) techniques, extraction of relevant information from free-text clinical notes in Electronic Medical records is often expensive and time-consuming [14]. Traditional approaches in NLP involve the construction of models based on expert-annotated corpora. These methods require extensive input from domain experts who have limited opportunity to review and provide feedback on the resulting models. Interactive Natural Language Processing holds promise in addressing this gap, towards improving clinical care as well as furthering clinical research faster (Section 2.4.1).

In my dissertation, I demonstrated the successful use of interactive NLP prototypes by clinicians for two example applications. I explore ideas from interactive machine learning (Section 2.1) by designing interface components to support *review*, *feedback* and *retrain* steps of an interactive NLP cycle. These systems allowed clinicians to build useful models with little or no initial training. In Chapter 4, NLPReViz served as an example of how clinicians could train their own models for retrospective research. We conducted user-studies with clinicians to evaluate our system and gather feedback for future re-design of similar systems. Next, I extended this approach for its application in a clinical care environment (Chapter 5).

We built an intelligent signout tool to help clinicians identify incidental findings. Similar to NLPReViz, we conducted a user-study to evaluate our tool. Apart from measuring model performance and usability scores, we included extrinsic evaluation metrics such as measuring task completion times, number of user actions needed, etc. Such evaluations support our hypothesis of using AI to augment clinical workflows. NLPReViz was designed for clinical researchers as target users, who are vested in the task of building global NLP models. In comparison, the users of the Intelligent Signouts tools are clinicians primarily engaged in care of individual patients and are building (local) NLP models. Together they serve as example tasks where interactive NLP can be adopted for analyzing clinical text, covering both clinical care and research applications. A real-world application may adapt a combination of ideas from both these example use-cases. These applications have different design requirements– for example, an intelligent application could support clinical workflows ('Intelligent Signout'-like use case), but also allow periodic inspection of the NLP models by senior team members or administrators (NLPReViz-like use case).

Further, while NLPReViz deals with the task of binary classification at the document level, Intelligent Signouts extends the learning problem for identifying relevant text spans within a full-text patient note. This can be seen as a step beyond a simple classification task – where we are not only classifying whether a document is important but are also interested in identifying relevant portions within the important document as well. Future applications of interactive NLP may tackle harder NLP tasks such as named-entity recognition, identifying relations, time-series analysis, natural language understanding and building question answering systems (Figure 4). These initial prototypes described in this dissertation help us understand the design of interactive NLP tools for such wider clinical applications. In both prototypes, we used simple machine learning pipelines for faster speed and easier implementations to demonstrate our interactive approach.

For evaluation of both the prototype tools, we did not integrate our approach with real EMR systems. For the incidentals problem, for example, we disabled the NLP predictions in the control condition of our prototype to simulate the traditional workflow for identifying incidentals. While we received very positive feedback for our tools and some participants even commented on how they preferred using these prototypes over the existing EMRs, our tools dealt with very specific user-study tasks. Integration with the EMR is a significantly more challenging problem. Considering the incidentals problem again as an example, physicians may review note notes for many different objectives at once in the real world, and not for singular tasks such as identifying incidentals. When we move from 'lab' user studies to 'field' or 'problem impact' studies [96], we require deeper explorations, such as long term case studies, that can influence design decisions.

Future work in NLP modeling may involve an exploration of more competitive classification performance as well as using better feature representation methods, such as our work on modeling incidentals [93]). Future explorations can include models for incorporating semi-supervised learning, positive and unlabeled classes (eg. identified vs. missed incidentals, and specifying "irrelevant" spans in NLPReViz), building collaborative models for a team, using soft-labels based on user-experience, handling noisy labels, and active learning.

Building models interactively requires establishing guidelines about how humans and AI algorithms should interact and collaborate [100, 101]. These principles will require a systematic study of prototype systems for specific applications and target users. This will lead to more opportunities for future research at all three steps of the interactive learning cycle:

1. **Review**
   (i) Make black-box models *explainable* and *interpretable*. We have some initial work in clinical informatics for making machine learning on EMR data more interpretable [91, 90], however, there is scope for Introducing transparency in modeling allows the users to have confidence in them. Other open problems include defining measures of confidence, transparency etc.

   (ii) *Safety* is another critical area for research. One of the issues studied in great details is the problem of *alert fatigue* [102, 103]. One may need to make choices between easily dismissible highlights vs. critical alerts, for example, for building interactive systems. Next, we also need to look into minimizing *automation bias* and preventing clinicians from becoming complacent due to automation.

2. **Feedback**

   (i) Allow *richer feedback* and combine data from *multiple modalities*. In my dissertation, I focused only on the free-text component of the EMR. Future work may delve in combining multiple modalities such as text plus images, and building combined models. This would require an exploration into novel ways of presenting data and providing feedback.

   (ii) Models that can be built in *collaboration* by a team. This would require designing systems that can manage user roles, modeling expertise and also manage conflicts.

   (iii) Care practices may be constantly updated and revised. As a result, interactive learning systems should able to handle *evolving guidelines* over a period of time for the same concept.

   (iv) Implement better *active learning* strategies for minimizing feedback costs [28].

3. **Re-train**

   (i) Another unsolved problem is in *estimating model performance* in the absence of a labeled gold standard in many real-world tasks. In both of the proposed tools, we did little to address the problem of letting the users know when the model is good enough. This could be done based on how the models perform against a held-out set and/or examples already reviewed and labeled manually.

   (ii) Explain model changes (live re-training vs. overnight updates) and *performance metrics*. Visualization of model performance metric over time could help address this need.

   (iii) Build *continuously learning* systems. Allow users to revert changes and switch between model revisions [37, 59].

In this dissertation, I presented two example applications with clinicians as end-users. While some of these principles may be extended for general Human-AI collaborations tasks (such the use of NLPReViz to classify legal text [104]), other applications will require a narrower focus on the target users– *clinicians*, *researchers*, and also *patients* in identifying and supporting these AI collaborations. The history of Biomedical Informatics deals with the conversion of medical knowledge into a computable form. Newer machine learning techniques have reinvigorated this possibility in continuously learning, intelligent EMR systems. The biggest obstacle in building such systems that can learn from EMR data is the lack of training labels. “Human-in-the-loop” and interactive methods reduce the need for labeled examples upfront and bring machine learning closer to end-users who consume these models. With the continuously learning interactive learning approaches as well as advances in unsupervised machine learning, not only we have the potential to support end-users of these models, but also contribute completely new insights to medical knowledge [105].

BIBLIOGRAPHY

[1] V. Nair, M. Kaduskar, P. Bhaskaran, S. Bhaumik, and Hodong Lee. Preserving narratives in electronic health records. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pages 418–421, Nov 2011.

[2] Marc Berg, Chris Langenberg, Ignas v.d Berg, and Jan Kwakkernaat. Considerations for sociotechnical design: experiences with an electronic patient record in a clinical context. *International Journal of Medical Informatics*, 52(1):243 – 251, 1998.

[3] A. N. Kho, L. V. Rasmussen, J. J. Connolly, P. L. Peissig, J. Starren, H. Hakonarson, and M. G. Hayes. Practical challenges in integrating genomic data into the electronic health record. *Genet. Med.*, 15(10):772–778, Oct 2013.

[4] Kasper Jensen, Cristina Soguero-Ruiz, Karl Oyvind Mikalsen, Rolv-Ole Lindsetmo, Irene Kouskoumvekaki, Mark Girolami, Stein Olav Skrovseth, and Knut Magne Augestad. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports*, 7:46226 EP –, Apr 2017. Article.

[5] American Medical Association. AMA calls for design overhaul of electronic health records to improve usability. *Press Release*, Sep 2014.

[6] Charlene R Weir and Jonathan R Nebeker. Critical issues in an electronic documentation system. *AMIA Annual Symposium Proceedings*, 2007:786–790, 2007.

[7] M. C. Wright, S. Dunbar, B. C. Macpherson, E. W. Moretti, G. Del Fiol, J. Bolte, J. M. Taekman, and N. Segall. Toward Designing Information Display to Support Critical Care. A Qualitative Contextual Evaluation and Visioning Effort. *Appl Clin Inform*, 7(4):912–929, Oct 2016.

[8] D.M. Zulman, N.H. Shah, and A Verghese. Evolutionary pressures on the electronic health record: Caring for complexity. *JAMA*, 2016.

[9] Stephen A Martin and Christine A Sinsky. The map is not the territory: medical records and 21st century practice. *The Lancet*, 388(10055):2053–2056, 05 2017.

[10] Orit Manor-Shulman, Joseph Beyene, Helena Frndova, and Christopher S. Parshuram. Quantifying the volume of documented clinical information in critical illness. *Journal of Critical Care*, 23(2):245 – 250, 2008.

[11] K. A. Artis, E. Dyer, V. Mohan, and J. A. Gold. Accuracy of Laboratory Data Communication on ICU Daily Rounds Using an Electronic Health Record. *Crit. Care Med.*, 45(2):179–186, Feb 2017.

[12] S. Malmasi, N. L. Sandor, N. Hosomura, M. Goldberg, S. Skentzos, and A. Turchin. Canary: An NLP Platform for Clinicians and Researchers. *Appl Clin Inform*, 8(2):447–453, May 2017.

[13] C Friedman and S B Johnson. Natural Language Processing in Biomedicine. In E H Shortliffe and J J Cimino, editors, *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, Health Informatics, chapter 8, pages 312–343. Springer, 2006.

[14] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543, 2011.

[15] Richard S. Dick, Elaine B. Steen, Don E. Detmer, and Institute of Medicine (U.S.). Committee on Improving the Patient Record. *The computer-based patient record: an essential technology for health care*. National Academy Press, Washington, D.C, 1997.

[16] L. Ohno-Machado. Realizing the full potential of electronic health records: the role of natural language processing. *J Am Med Inform Assoc*, 18(5):539, 2011.

[17] Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Anthony Gitter, and Casey S. Greene. Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv*, 2017.

[18] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M. Pintea, and Vasile Palade. Towards interactive machine learning (iml): Applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In Francesco Buccafurri, Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, editors, *Availability, Reliability, and Security in Information Systems*, pages 81–95, Cham, 2016. Springer International Publishing.

[19] Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.

[20] Malcolm Ware, Eibe Frank, Geoff Holmes, Mark A. Hall, and Ian H. Witten. Interactive machine learning: letting users build classifiers. *Int. J. Hum.-Comput. Stud.*, 55:281–292, 2001.

[21] J.A. Fails and D.R. Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45. ACM, 2003.

[22] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4):105–120, 2014.

[23] Saleema Amershi, Bongshin Lee, Ashish Kapoor, Ratul Mahajan, and Blaine Christian. Cuet: Human-guided fast and accurate network alarm triage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 157–166, New York, NY, USA, 2011. ACM.

[24] Jesse Vig, Shilad Sen, and John Riedl. Navigating the tag genome. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 93–102. ACM, 2011.

[25] Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 147–156, New York, NY, USA, 2011. ACM.

[26] Duen Horng Chau, Aniket Kittur, Jason I. Hong, and Christos Faloutsos. Apolo: Making sense of large network data by combining rich user interaction and machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 167–176, New York, NY, USA, 2011. ACM.

[27] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2839–2848, 2012.

[28] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[29] Richard S. Sutton and Andrew G. Barto. Reinforcement learning - an introduction. In *Adaptive computation and machine learning*, 1998.

[30] Christopher Bishop. *Linear Models for Regression*, chapter 3. Springer-Verlag, New York, 2006.

[31] Jürgen Bernard, Matthias Zeppelzauer, Michael Sedlmair, and Wolfgang Aigner. A Unified Process for Visual-Interactive Labeling. In Michael Sedlmair and Christian Tominski, editors, *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2017.

[32] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[33] Robert Spence. *Information Visualization: Design for Interaction*. Prentice Hall, 2007.

[34] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. Effective End-User Interaction with Machine Learning. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 1529–1532, 2011.

[35] *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Bethesda, MD, 1978.

[36] Nina Hallowell. Research or clinical care: what's the difference? *Journal of Medical Ethics*, 44(6):359–360, 2018.

[37] Todd Kulesza, Margaret Burnett, Weng-keen Wong, and Simone Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, pages 126–137, 2015.

[38] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. Cueflik: Interactive concept learning in image search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 29–38, New York, NY, USA, 2008. ACM.

[39] M. Cakmak and A. L. Thomaz. Optimality of human teachers for robot learners. In *2010 IEEE 9th International Conference on Development and Learning*, pages 64–69, Aug 2010.

[40] Jesse Vig, Shilad Sen, and John Riedl. Navigating the tag genome. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, pages 93–102, New York, NY, USA, 2011. ACM.

[41] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1343–1352, New York, NY, USA, 2010. ACM.

[42] Nadia Boukhelifa, Anastasia Bezerianos, and Evelyne Lutton. Evaluation of interactive machine learning systems. *CoRR*, abs/1801.07964, 2018.

[43] C. Breazeal. Role of expressive behaviour for robots that learn from people. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 364(1535):3527–3538, Dec 2009.

[44] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, pages 82–91, New York, NY, USA, 2007. ACM.

[45] Stephanie L Rosenthal and Anind K Dey. Towards maximizing the accuracy of human-labeled sensor data. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 259–268. ACM, 2010.

[46] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M. Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. Why-oriented end-user debugging of naive bayes text classification. *ACM Trans. Interact. Intell. Syst.*, 1(1):2:1–2:31, October 2011.

[47] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. Overview based example selection in end user interactive concept learning. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, pages 247–256. ACM, 2009.

[48] Gaurav Trivedi, Phuong Pham, Wendy W Chapman, Rebecca Hwa, Janyce Wiebe, and Harry Hochheiser. NLPReViz: an interactive tool for natural language processing on clinical text. *Journal of the American Medical Informatics Association*, 25(1):81–87, 2018.

[49] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S. Tan. Ensemblematrix: Interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1283–1292, New York, NY, USA, 2009. ACM.

[50] Yue Wang, Kai Zheng, Hua Xu, and Qiaozhu Mei. Interactive medical word sense disambiguation through informed learning. *Journal of the American Medical Informatics Association*, page ocy013, 2018.

[51] Glenn T Gobbel, Jennifer Garvin, Ruth Reeves, Robert M Cronin, Julia Heavirland, Jenifer Williams, Allison Weaver, Shrimalini Jayaramaraja, Dario Giuse, Theodore Speroff, Steven H Brown, Hua Xu, and Michael E Matheny. Assisted annotation of medical free text using raptat. *Journal of the American Medical Informatics Association*, 21(5):833–841, 2014.

[52] Glenn T. Gobbel, Ruth Reeves, Shrimalini Jayaramaraja, Dario Giuse, Theodore Speroff, Steven H. Brown, Peter L. Elkin, and Michael E. Matheny. Development and evaluation of raptat: A machine learning system for concept mapping of phrases from medical narratives. *Journal of Biomedical Informatics*, 48:54 – 65, 2014.

[53] L. W. D'Avolio, T. M. Nguyen, S. Goryachev, and L. D. Fiore. Automated concept-level information extraction to reduce the need for custom software and rules development. *Journal of the American Medical Informatics Association*, 18(5):607–613, 2011.

[54] Philip V. Ogren. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[55] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[56] Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. Clamp a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, page ocx132, 2017.

[57] Elijah Mayfield and Carolyn Penstein Rosé. Open source machine learning for text. *Handbook of automated essay evaluation: Current applications and new directions*, 2013.

[58] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[59] Jaegul Choo, Changhyun Lee, Chandan K. Reddy, and Haesun Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1992–2001, December 2013.

[60] Jason Chuang, Daniel Ramage, Christopher D. Manning, and Jeffrey Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *ACM Human Factors in Computing Systems (CHI)*, 2012.

[61] Spence Green, Jeffrey Heer, and Christopher D. Manning. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 439–448, New York, NY, USA, 2013. ACM.

[62] Gaurav Trivedi, Phuong Pham, Wendy Chapman, Rebecca Hwa, Janyce Wiebe, and Harry Hochheiser. An interactive tool for natural language processing on clinical text. In *4th Workshop on Visual Text Analytics (IUI TextVis 2015)*, Atlanta, Mar 2015.

[63] Gaurav Trivedi. Clinical text analysis using interactive natural language processing. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, IUI Companion '15, pages 113–116, New York, NY, USA, 2015. ACM.

[64] Omar F. Zaidan and Jason Eisner. Using "annotator rationales" to improve machine learning for text categorization. In *In NAACL-HLT*, pages 260–267, 2007.

[65] Martin Wattenberg and Fernanda B. Viegas. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, 2008.

[66] Ainur Yessenalina, Yisong Yue, and Claire Cardie. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1046–1056, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[67] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 3075–3084, New York, NY, USA, 2014. ACM.

[68] Henk Harkema, Wendy Webber Chapman, Melissa Saul, Evan S. Dellon, Robert E. Schoen, and Ateev Mehrotra. Developing a natural language processing application for measuring the quality of colonoscopy procedures. *JAMIA*, 18(Supplement):150–156, 2011.

[69] J. Brooke. SUS: a quick and dirty usability scale. In P. W. Jordan, B. Weerdmeester, A. Thomas, and I. L. Mclelland, editors, *Usability evaluation in industry*. Taylor and Francis, London, 1996.

[70] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[71] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*, pages 364–371. Elsevier, 2003.

[72] SK Mueller, C Yoon, and JL Schnipper. Association of a web-based handoff tool with rates of medical errors. *JAMA Internal Medicine*, 2016.

[73] Max V. Wohlauer, Kyle O. Rove, Thomas J. Pshak, Christopher D. Raeburn, Ernest E. Moore, Chad Chenoweth, Apoorva Srivastava, Jonathan Pell, Randall B. Meacham, and Mark R. Nehler. The computerized rounding report: Implementation of a model system to support transitions of care. *Journal of Surgical Research*, 172(1):11 – 17, 2012.

[74] L. D. Sanchez, D. T. Chiu, L. Nathanson, S. Horng, R. E. Wolfe, M. L. Zeidel, K. Boyd, C. Tibbles, S. Calder, J. Dufresne, and J. J. Yang. A Model for Electronic Handoff Between the Emergency Department and Inpatient Units. *J Emerg Med*, 53(1):142–150, Jul 2017.

[75] Marc-David Munk, Andrew B. Peitzman, David P. Hostler, and Allan B. Wolfson. Frequency and follow-up of incidental findings on trauma computed tomography scans:

Experience at a level one trauma center. *The Journal of Emergency Medicine*, 38(3):346 – 350, 2010.

[76] J. L. Sperry, M. S. Massaro, R. D. Collage, D. H. Nicholas, R. M. Forsythe, G. A. Watson, G. T. Marshall, L. H. Alarcon, T. R. Billiar, and A. B. Peitzman. Incidental radiographic findings after injury: dedicated attention results in improved capture, documentation, and management. *Surgery*, 148(4):618–624, Oct 2010.

[77] A. Salim, B. Sangthong, M. Martin, C. Brown, D. Plurad, and D. Demetriades. Whole body imaging in blunt multisystem trauma patients without obvious signs of injury: results of a prospective study. *Arch Surg*, 141(5):468–473, May 2006.

[78] B Lumbreras, L Donat, and I Hernndez-Aguado. Incidental findings in imaging diagnostic tests: a systematic review. *The British Journal of Radiology*, 83(988):276–289, 2010. PMID: 20335439.

[79] M. K. James, M. P. Francois, G. Yoeli, G. K. Doughlin, and S. W. Lee. Incidental findings in blunt trauma patients: prevalence, follow-up documentation, and risk factors. *Emerg Radiol*, 24(4):347–353, Aug 2017.

[80] Meliha Yetisgen-Yildiz, Martin L Gunn, Fei Xia, and Thomas H Payne. Automatic identification of critical follow-up recommendation sentences in radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1593. American Medical Informatics Association, 2011.

[81] John Zech, Margaret Pain, Joseph Titano, Marcus Badgeley, Javin Schefflein, Andres Su, Anthony Costa, Joshua Bederson, Joseph Lehar, and Eric Karl Oermann. Natural languagebased machine learning models for the annotation of clinical radiology reports. *Radiology*, 0(0):171093, 0. PMID: 29381109.

[82] Meliha Yetisgen, Prescott Klassen, Lucas McCarthy, Elena Pellicer, Tom Payne, and Martin Gunn. Annotation of clinically important follow-up recommendations in radiology reports. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 50–54, 2015.

[83] R. Pivovarov and N. Elhadad. Automated methods for the summarization of electronic health records. *J Am Med Inform Assoc*, 22(5):938–947, Sep 2015.

[84] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. *Mining text data*, pages 43–76, 2012.

[85] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 68–73, New York, NY, USA, 1995. ACM.

[86] J. C. Denny, A. Spickard, K. B. Johnson, N. B. Peterson, J. F. Peterson, and R. A. Miller. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of American Medical Informatics Association*, 16(6):806–815, 2009.

[87] Ying Li, Sharon Lipsky Gorman, and Noémie Elhadad. Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 744–750, New York, NY, USA, 2010. ACM.

[88] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 213–220, New York, NY, USA, 2008. ACM.

[89] Yoon Kim. Convolutional neural networks for sentence classification, 2014.

[90] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks, 2017.

[91] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Peter J. Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Gavin E. Duggan, Gerardo Flores, Michaela Hardt, Jamie Irvine, Quoc Le, Kurt Litsch, Jake Marcus, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael Howell, Claire Cui, Greg Corrado, and Jeff Dean. Scalable and accurate deep learning for electronic health records, 2018.

[92] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[93] Gaurav Trivedi, Charmgil Hong, Esmaeel R. Dadashzadeh, Robert M. Handzel, Harry Hochheiser, and Shyam Visweswaran. Identifying incidental findings from radiology reports of trauma patients: An evaluation of automated feature representation methods. *International Journal of Medical Informatics*, 129:81 – 87, 2019.

[94] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[95] Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems, CHI '11, pages 147–156, New York, NY, USA, 2011. ACM.

[96] Charles P. Friedman and Jeremy C. Wyatt. *Evaluation Methods in Biomedical Informatics (Health Informatics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[97] Jeff Sauro. *A practical guide to the system usability scale: background, benchmarks and best practices*. CreateSpace, Denver, CO, 2011. OCLC: 1078997064.

[98] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, Oct 2001.

[99] E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.*, 25(1):44–56, Jan 2019.

[100] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '19. ACM, 2019.

[101] Jeffrey Heer. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6):1844–1850, 2019.

[102] Aaron S Kesselheim, Kathrin Cresswell, Shobha Phansalkar, David W Bates, and Aziz Sheikh. Clinical decision support systems could be modified to reduce alert fatiguewhile still minimizing the risk of litigation. *Health affairs*, 30(12):2310–2317, 2011.

[103] Aaron S Kesselheim, Kathrin Cresswell, Shobha Phansalkar, David W Bates, and Aziz Sheikh. Clinical decision support systems could be modified to reduce alert fatiguewhile still minimizing the risk of litigation. *Health affairs*, 30(12):2310–2317, 2011.

[104] Jaromir Savelka, Gaurav Trivedi, and Kevin Ashley. Applying an interactive machine learning approach to statutory analysis. In *JURIX 2015 - the 28th International Conference on Legal Knowledge and Information Systems*, Dec 2015.

[105] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.