Bridging the Natural Language Processing Gap: An Interactive Clinical Text Review Tool

Gaurav Trivedi[1], Phuong Pham[1], Wendy Chapman[2], Rebecca Hwa[1], Janyce Wiebe[1], Harry Hochheiser[1]
[1]University of Pittsburgh, Pittsburgh, PA; [2]University of Utah, Salt Lake City, UT
trivedigaurav@pitt.edu

Barriers to NLP Adoption

- We have a long history of resarch on NLP methods in the clinical domain [1].
- However, the complexity of unstructured clinical text makes analysis a hard problem and its accuracy varies.
- Domain experts may be able to fix problems with the models but they may not be familiar with symbolic and machine learning techniques.

Our Solution: An Interactive Tool for NLP on Clinical Text

Our goal is to close the NLP gap by providing clinical researchers with highly-usable tools that will facilitate the process of reviewing NLP output, identifying errors in model prediction, and providing feedback that can be used to retrain or extend models to make them more effective.

We have developed an interactive web-based tool that facilitates both the review of binary variables extracted from clinical records, and the provision of feedback that can be used to improve the accuracy of NLP models.

Design Requirements

We have built upon ideas in *Visualization*, *Interactive Machine Learning* and *Interface Design* research.

Our design requirements are summarized as follows:

R1: The tool should make it easier for machine learning non-experts to work with NLP models.

R2: It should incorporate efficient mechanisms for annotation and labeling, and also encourage feedback that is consistent and informative.

R3: The interactive components should support the entire interactive machine learning loop - i.e. a *review*, *feedback* and *retrain* cycle.

Acknowledgments

Our demo uses an example dataset of colonoscopy reports and is based on the work done by Harkema et. al. [2]. This research is supported by NIH grant 5R01LM010964.

References

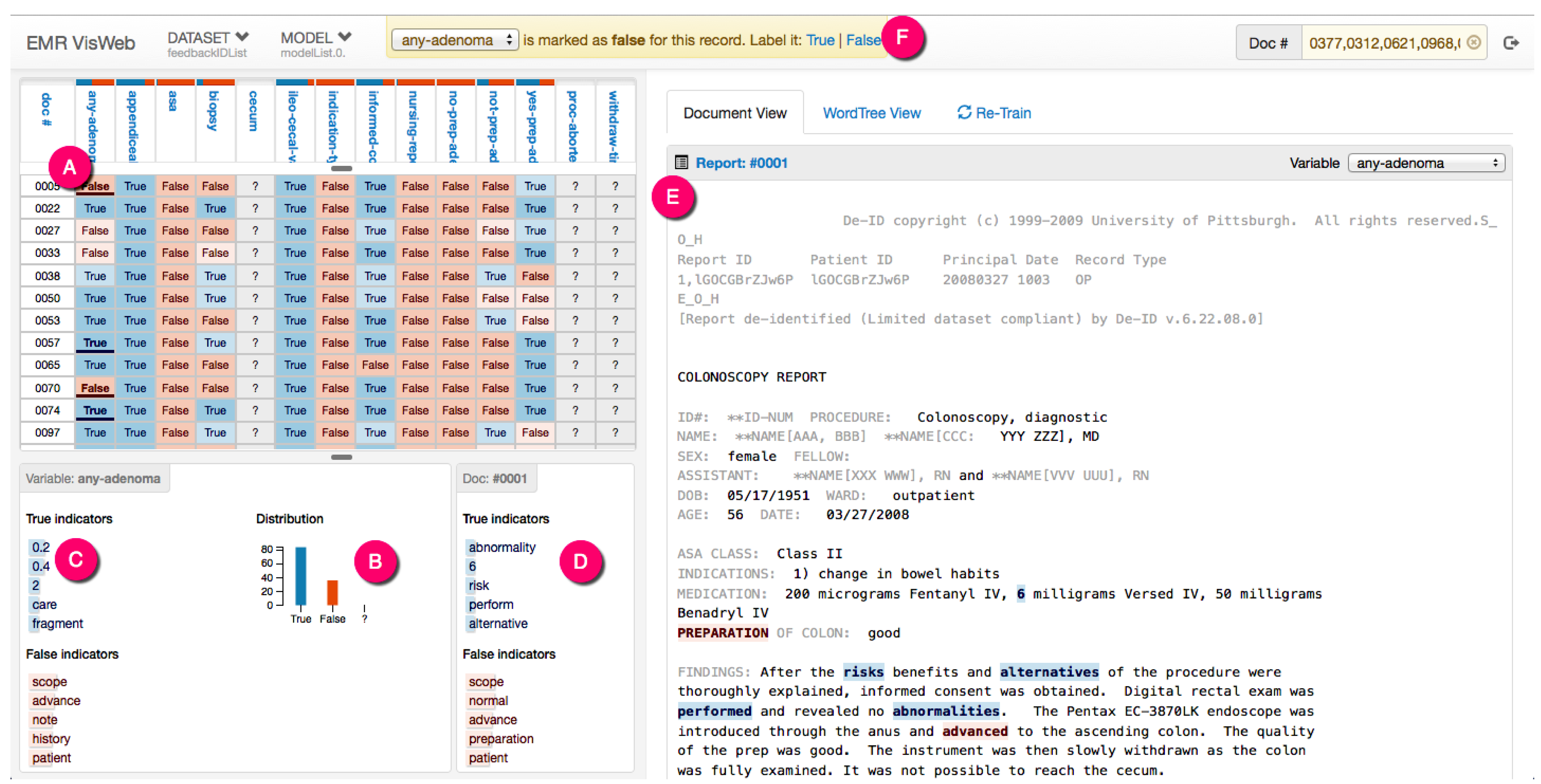
[1] Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., and Uzuner, O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association 18*, 5 (2011), 540–543.

[2] Harkema, H., Chapman, W. W., Saul, M., Dellon, E. S., Schoen, R. E., and Mehrotra, A. Developing a natural language processing application for measuring the quality of colonoscopy procedures. *JAMIA 18*, Supplement (2011), 150–156.

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

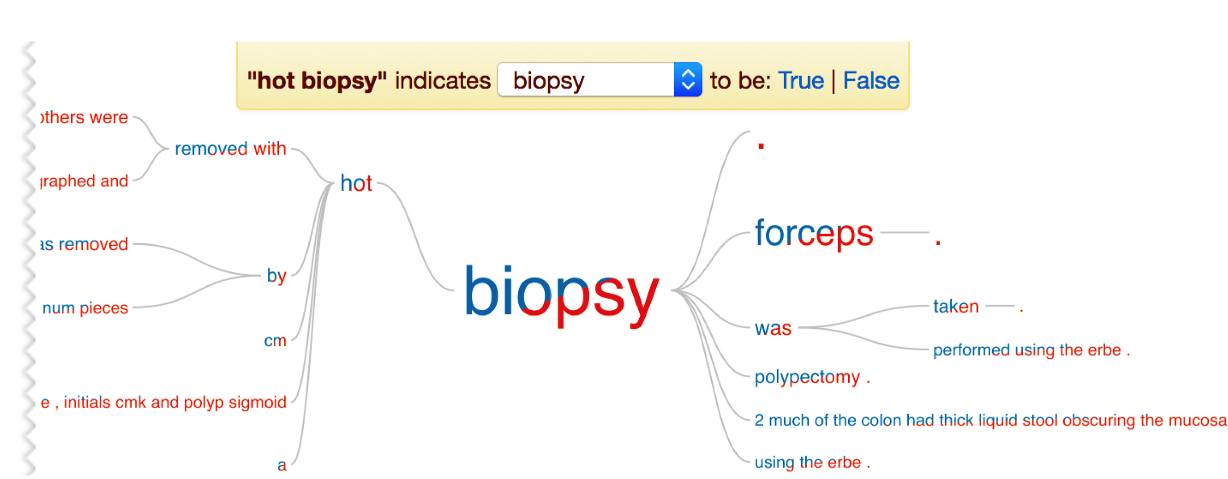
[4] Brooke, J. SUS: a quick and dirty usability scale. In *Usability evaluation in industry*, P. W. Jordan, B. Weerdmeester, A. Thomas, and I. L. McLelland, Eds. Taylor and Francis, London, 1996.

Interface Design

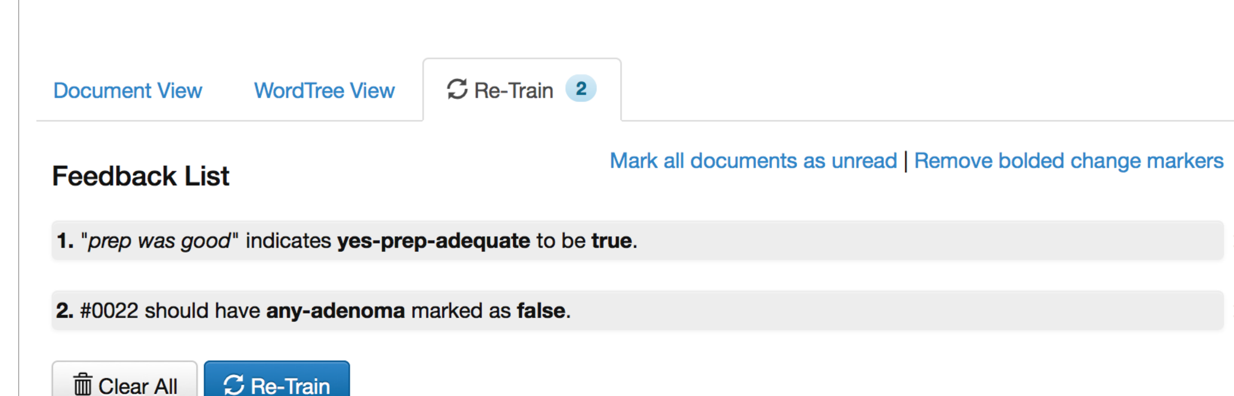


A The **Grid view** shows the extracted variables in columns and individual documents in rows, providing an overview of NLP results. Below the grid, we have statistics about the active variable with B the distribution of the classifications and C the list of top indicators aggregated across all the documents 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 the right-click context menu.

The **WordTree** [3] **view** provides the ability to search for and explore word sequence patterns found across the documents, and to provide feedback for retraining.



The **Re-Train view** lists user-provided feedback, including any potential inconsistencies, and specifies changes in variable assignments due to retraining.



A demo video of the tool is available at `http://vimeo.com/trivedigaurav/emr-demo`.

User Study and Results

We conducted a formative user study with five clinicians and clinical researchers as participants to gain insight into usability factors of the tool that may be associated with errors or confusion, and to identify opportunities for improvement via re-design or implementation of new functionality.

We used the **System Usability Scale** [4] consisting of 10-questions on a 5-point Likert scale to help get a global view of subjective assessments of usability. The average SUS score was 70.5 out of 100.

A summary of recommendations inferred from the user study for is given below:

Category	Recommendation
Workflow	1. Allow sorting (or filtering) of the documents in the grid based on the prediction probabilities. This would make it easier for the users to prioritize documents to review.
	2. Add a button to open the next-in-line document for review. The order may be decided either trivially based on ID number or by using an active learning approach. This would save the users to navigate through the grid when they don't have their own strategy for selecting documents for review.
WordTree	1. Change the layout of the tool to show the WordTree view along with the document view. This would allow the user to quickly go through the full report text when the wordtree tree is unable to provide sufficient contextual information.
	2. Allow selection of multiple branches in the tree to give feedback on multiple paths in the tree at once.
Feedback	1. Provide a feedback mechanism to specify that a text span does not indicate either of the classes. This would allow the user to remove non-informative but possibly misleading features in re-training.
Re-Training	1. Perform auto-retraining in the background when a sufficient number of feedback items have been provided by the user.
	2. Provide a built-in mechanism to validate and generate a performance report for the current model against a held-out test set.

Future efforts will involve incorporating these recommendations and conducting an empirical evaluation.