Project Summary

CHS: Small: Interactive Machine Learning for Text Analysis

Jeffrey Heer
University of Washington

Text – including original documents, online correspondence and transcribed speech – is a fundamental data type in a variety of research domains. Tasks requiring text analysis include identifying medical terms in research papers or patient-authored text; finding linguistic markers of affect, politeness or leadership in online discourse; tracking policies across pieces of legislation; and determining consumer sentiments about a product from social media. Across these examples, analysis involves the *recognition* and/or *classification* of phrases or textual categories: researchers iteratively develop or use pre-existing labeling schemas; annotate terms, sentences or full documents; and train and apply statistical classifiers to analyze data at scale.

As automated text mining approaches improve, the *process* of text analysis remains dominated by human effort and supervision. Researchers must collect and manage large text collections, select or develop coding schemes, annotate a subset of the data (either directly or by training coders), identify predictive textual features, tune algorithm parameters, and assess the results of applying automated methods to the full dataset. This process does not proceed in a linear fashion, instead requiring iteration within and across phases, often switching among tools in a manner that stymies provenance tracking and replication.

We propose to develop an integrated, interactive software system to support the process of classification-oriented text analysis. We hypothesize that novel interfaces and supporting algorithms can reduce time and effort and make text analysis methods more accessible to researchers, while retaining – and likely improving – the quality of the resulting classifiers. We will develop an end-to-end system that includes management of documents and metadata; a visual interface for integrated and iterative schema generation, text annotation and model evaluation; and a runtime for managing and comparing multiple learned classifiers.

Core intellectual challenges include the design and evaluation of visual analysis and interactive machine learning techniques, which enable domain experts who may lack training in statistical machine learning to effectively analyze text data. We envision a "virtuous cycle" in which analysts formulate schemas and provide annotations, visualizations facilitate understanding of data and models, and automated methods generalize user input and suggest additional data and features for annotation. We aim to help users also track their progress and replicate analyses. We hope to significantly enhance existing practices of text analysis.

Intellectual Merit: Our research will develop new technical contributions and experimental results. On the technical front, we will investigate system architectures for mixed-initiative text classification; novel user interfaces and visualizations for annotation and model evaluation; interactive techniques for improved feature selection; active learning methods for adaptive sampling of instances and features to label; and facilities for collaborative and crowdsourced labeling. We will conduct evaluations with domain scientists and crowdsourced workers to assess how our methods affect the time and effort required for text analysis, the quality of the resulting classifiers, and the potential biases introduced by automated methods.

Broader Impacts: Our work will lower barriers to entry and enable faster and higher-quality text analysis. The resulting tools can positively impact disciplines that analyze text data. We will work hand-in-hand with our collaborators in multiple domains (health & addiction studies, political science, psychology, sociology) to substantiate these benefits. We will share our tools as open source software runnable as a web service, and leverage our software platform in classroom teaching and undergraduate research.

Keywords: text analysis; data visualization; interactive machine learning; active learning; crowdsourcing.