Using Machine Learning and Visualization for Qualitative Inductive Analyses of Big Data

Harshini Priya Muthukrishnan*
University of Colorado Boulder

Danielle Albers Szafir†
University of Colorado Boulder

ABSTRACT

Many domains require analyst expertise to determine what patterns and data are interesting in a corpus. However, most analytics tools attempt to prequalify "interestingness" using algorithmic approaches to provide exploratory overviews. This overview-driven workflow precludes the use of qualitative analysis methodologies in large datasets. This paper discusses a preliminary visual analytics approach demonstrating how visual analytics tools can instead enable expert-driven qualitative analyses at scale by supporting computer-in-the-loop mixed initiative approaches. We argue that visual analytics tools can support rich qualitative inference by using machine learning methods to continually model and refine what features correlate to an analyst's on-going qualitative observations and by providing transparency into these features in order to aid analysts in navigating large corpora during qualitative analyses. We illustrate these ideas through an example from social media analysis and discuss open opportunities for designing visualizations that support qualitative inference through computer-in-the-loop approaches.

Index Terms: Qualitative analysis—Qualitative inductive methods—inductive analysis—computer-in-the-loop; Human-centered computing—Visualization—Visualization techniques

1 INTRODUCTION

Analysts typically explore a large text corpus by first processing the data using algorithms like topic modeling and visualizing the algorithm's outputs. While the use of automated algorithmic approaches can help quickly distill large-scale patterns, their focus on corpus-scale trends may cause them to miss capturing sparse yet significant patterns in a given corpus. When dealing with big data, as long as we are looking to identify dominant patterns or perform a search for a specific pattern, automated methods do well. However, many interesting problems in data lie in the small-scale patterns or those not well defined *a priori*. For example, experts can use their own intuitions to identify patterns of emotional distress indicative of suicide or depression in social media data. Emotional distress makes up a small percentage of social media posts and is not dominant enough for traditional automated techniques to capture.While qualitative expert intuitions often drive the detection of these small-scale phenomena, these patterns often correlate with quantifiable statistical features such as first person pronoun usage [2]. This correlation suggests that automated approaches may be able to detect these patterns from qualitative observations if we can model what an analyst finds important.

We argue for automated approaches that allow analysts to work collaboratively with the algorithm and let them drive the exploration starting from details and moving to an overview in order to analyze data based on qualitative expert observations. While most visualizations use a deductive pipeline following an overview first and details on demand workflow, our goal is to create an inductive analysis pipeline: we want to allow analysts to drive the analysis process to build up new theories beginning from specific details within the data [18]. This design of an inductive visualization pipeline follows directly from techniques used by qualitative analysts engaged in inductive methods that progressively organize data into increasingly abstract units of information, thereby building patterns and themes from the bottom up—from details to overview.

Our approach provides a mixed-initiative paradigm that allows automated systems to interact with analysts in two ways, as an annotator that suggests similarly occurring instances in the corpus and as a collaborator that can identify and iterate predictions made by the model. We embody this method in a preliminary prototype called QualVis which uses machine learning to model qualitative observations provided by an analyst to interactively support techniques from qualitative inductive methods to discover patterns in large corpora. QualVis showcases this functionality in the context of the identification of emotional distress in social media by modeling a qualitative analysts' natural workflow of finding patterns in a text corpus, as described in and using data from Brubaker et al. [2]. Emotional distress can be defined as an expression of grief and distress following the occurrence of a major life event. In this paper, we study the emotional distress expressed by people on social networking sites following the death of loved ones. Emotional distress is not well detected by traditional algorithms; however, trained analysts can detect emotional distress through manual annotation, inductively building this inference from a thoughtful investigation of a corpus in context. Our collaborators have reflected on statistical patterns in an inductive labeling of emotional distress [2], making it an ideal test case for exploring how mixed initiative systems might support inductive inference.

In order to automate pattern discovery and enable user-initiated data exploration, we use supervised ML algorithms that fit a given labeled training data (emotional distress or not emotional distress) to expert-labeled data in order to seed our system. We then use these models to provide interactive guidance that helps analysts find other relevant passages throughout the corpus. As analysts read through a text, they can highlight relevant passages of interest. For each passage, the system identifies a subset of the corpus that is similar to the analyst's chosen passage. The system recommends similar sentences from the corpus using a weighted metric combining cosine similarity and classification probabilities from the emotional distress model. By narrowing the analysts' focus to a subset of the original corpus, we scale the identification of patterns to large corpora which would have otherwise been a human resource intensive process.

Interactive visualizations allow analysts to reason about the recommended passages, lending transparency to the model and increasing agency and collaboration. Our system augments a text annotation interface with two additional visualizations. The first shows data pertaining to the similarity detection and classification models by conveying the reasons why the model has given the results it has currently made. The second visualization provides a glyph-based scatterplot that shows an overview of all annotated documents with respect to learned features and helps the analyst identify patterns across the corpus. Our visualizations help analysts explore a corpus across multiple levels of detail starting from raw text of a document

*e-mail: harshini.muthukrishnan@colorado.edu
†e-mail: danielle.szafir@colorado.edu

Table 1: Set of features used to train the Naive Bayes model

Feature	Source of the feature
Count of First person singular pronouns	Presence of the words "i, me, my, mine"
Count of Negations	Presence of the words "not, n't, never, neither, nobody, no, none, nor, nothing, nowhere"
Count of Second person plural pronouns	Presence of the words "you, your, yours, he, she, it, him, her, its, his, hers, we, us, our, ours, they, them, their, theirs"
Count of Past tense verbs	Presence of Parts of speech like 'VBD', 'VBN' as defined by the pos tagger from nltk package
Count of Future tense verbs	Presence of Parts of speech like 'MD' as defined by the pos tagger from nltk package
Count of Adverbs	Presence of Parts of speech like 'RR', 'RBR', 'RBS', 'WRB' as defined by the pos tagger from nltk package
Count of Prepositions	Presence of Parts of speech like 'IN' as defined by the pos tagger from nltk package
Count of Conjunctions	Presence of Parts of speech like 'CC' as defined by the pos tagger from nltk package
Emotion	Presence of emotion intensity equal to 100 as defined by the senpy library
Length	Number of words on tokenization using the nltk word tokenizer

to related passages in other documents by using specific passages of interest to drive these representations. The increasing abstractions of visualization play a significant role for user interaction with the system due to the analysts' increasing hierarchal perception when scanning big corpora during inductive qualitative analysis processes. This work represents preliminary steps towards understanding how visualizations might support expert-driven, inductive analysis workflows to expand the kinds of insights visualization tools enable.

2 Related Work

Qualitative analysis allows analysts to leverage their own expertise and observations to build theory from data. Qualitative analysis can be done using a number of different techniques like thematic analysis and grounded theory to generate insights. Open qualitative coding is one of the most widely used qualitative approaches to inductively label and categorize emergent concepts while maintaining theoretical freedom [17]. This technique is especially useful for unstructured big data like social media [1]. Preliminary visualization systems like that from Chadrasegaran et al. [3] leverage common NLP techniques (part of speech tagging and topic modeling) to provide overview visualizations to support grounded theory analyses through top-down open coding.

Qualitative inductive methods (QIMs) allow analysts to iteratively generate a theory by instead building up from specific examples to general ideas [18]. Analysts first review raw text to find interesting exemplars. They then build links between these exemplars, and finally generate new theories from the linked data. At present, qualitative inductive methods are usually done manually using paper and highlighter or naive general-purpose software like word processors and spreadsheets thus making the process extremely laborious. QIM-specific tools like ATLAS.ti[1], Dedoose[2] allow analysts to highlight text and assign codes. MAXQDA[3] goes a step further and supports basic descriptive and inferential statistics through the addition of descriptive attributes to the data. However, these tools still rely on analysts to navigate and synthesize exemplars in the corpus, providing little support for large-scale computational analysis or automated guidance as the analyst navigates their data. Our discussions with qualitative analysts suggest that they primarily deal with corpus scale by only analyzing a random subset of their data.

Our approach leverages human-in-the-loop methods to scale QIMs by mining user interactions (i.e., highlighting and code application). Human-in-the-loop approaches rely on the user to help to provide feedback to classification models that are used for labeling data samples. Active Learning is a popular human-in-the-loop approach where a classifier iteratively requests new data from a human annotator—known as an *oracle*—by posing queries to the oracle [16]. The queries typically ask the user to provide labels for data that is currently unlabeled. We propose a bottom-up version of active learning which does not require an *a priori* target research question and is therefore oracle-initiated feedback rather than classifier-initiated. Techniques like ELA [5] provide similar support for organically identifying labels, but require analysts to do so top-down, starting from a data overview. Our method instead lets analysts generate labels on the fly using qualitative codes and progressively learns these codes based on the exemplars provided by the analyst. The codes can be preemptively seeded using priors from related analyses. By reversing who initiates the labeling, we let the oracle decide the relevance of predictions and steer the refinement of the model for future predictions. Since qualitative analysis relies heavily on human expertise to discover interesting patterns in data, our approach builds on the concept of Active Learning where the user has the ability to drive the analysis task as opposed to a machine in human-in-the-loop approaches.

Automated annotation support tools like ALIA [4] help steer analysts towards interesting regions of a corpus using automated methods. However, our discussions with qualitative analysts suggest that oracle-initiated feedback requires that analysts can fluidly understand not just that a passage is similar, but also *why* it is similar to maintain agency and ensure that the analyst's expertise is driving the analysis rather than potentially spurious correlations.

One way to provide analysts with this insight is to use visualizations that explain the reasons for a given classification. Most of the work in explanatory visualization focuses on providing interpretation and insights into black box machine learning models. Explanatory visualizations can substantially increase how well people understand (and subsequently improve) an ML model [10]. Machine learning visualizations for text data can focus on breaking down explanations to word level using sparklines and word clouds [6] to show relationships and statistical information. Kangasrääsiö et al. [8] model user feedback using regression to create a timeline chart of the feedback history showing accuracy inferences made by the model alongside user feedback and adjustments. Krause et al. [9] present a feature-based visualization that shows the influence of features on prediction results by providing an interactive partial dependence diagnostics along with support for tweaking feature values. Using human interpretable features as building blocks for the interactive visualization requires little knowledge of the details regarding the working of the model. This paradigm is optimal for many qualitative analysts who can more efficiently compare the semantics of data features to their own intuitions than to try to decompose more complex elements specific to the classification pipeline.

3 Workflow Analysis

Qualitative inductive methods (QIMs) help the analysts to identify themes in data by starting with specific instances and then steadily synthesizing more general patterns characterizing these instances. This approach solely relies on human experts and is common in methods like thematic analysis, grounded theory and contextual

[1] www.atlasti.com
[2] www.dedoose.com
[3] www.maxqda.com

inquiry. While QIMs bring deep qualitative inference and domain expertise into data they are limited for large datasets as they exclusively rely on people to manually identify and synthesize relevant data. We illustrate this constraint in QualVis using a subset of the dataset prelabeled for emotional distress from Brubaker et al. [2] which examined 2,213 post-mortem comments posted to the profiles of 652 MySpace users following their deaths. We use a prelabeled corpus in our proof of concept as a pseudo-Wizard-of-Oz dataset to both provide a scaffold of priors for design iteration with our collaborators and to allow a basis for comparison of how well our automated techniques capture this sparse, qualitative phenomenon.

Conversations with qualitative analysts suggest that there is no single method or workflow that completely characterizes inductive analyses. Instead, analysts move between coding interesting passages, building links between passages, and synthesizing theory from collections of linked passages [18]. This identify-and-link paradigm shares many similarities to interactive labeling approaches like ELA [5]; however, analysts wish to approach the text with no *a priori* biases in their exploratory analyses, making algorithmic overviews or machine-led human-in-the-loop approaches overly constrained. By mimicking the bottom up workflow of inductive analyses, we instead argue that a details-first approach [11] using oracle-initiated active learning (that is, active learning where the labels and query exemplars are driven by the analyst's interaction with the system) is necessary to support scalable QIM practices.

Our proposed approach leverages QIM practices and applies them to large corpora by means of machine learning to augment ongoing analyses. Analysts can inductively analyze text by highlighting and labeling interesting patterns in the data mimicking paper-based manual annotation practices used by our collaborators. Based on an analyst's selection, the algorithm updates the classifier for the target label and computes and suggests relevant similar sentences from the larger corpus. Analysts can then explore related instances which the tool suggests. With the use of machine learning algorithms to suggest related data points we enable the analysts to strategically navigate large corpora: the algorithm processes data based on the labels provided during the analysis process to help analysts fluidly mine specific exemplars. In addition, predictions from the machine learning model will help the analysts to understand the bias in their interpretation and thereby help with exploration of data that would have otherwise been ignored due to the large size of the corpus.

Our tool relies on NLP and supervised learning methods to inductively generate patterns of interestingness from data. Our approach deviates from the traditional human-in-the-loop techniques through active learning where the model uses deductive reasoning through an *a priori* target question. We instead couple classifiers with QIM practices and inductively generate themes from the data based on the analyst's interaction with the tool during an on-going analysis. We model our tool using feature representations at varying levels starting from words and n-grams and moving to sentence-level quantitative parameters like word count and term co-occurrence (Table 1). This method aligns the tool with the traditional QIM practices and also helps with scaling for big data by leveraging automation to provide intelligent guidance for new passages of interest inferred from the analyst's interactions.

Current work practices with QIMs are often collaborative between different people. While the machine learning algorithms guide the analyst throughout the process with their predictions, they are not always capable of adapting to new data or finding new and emerging patterns. Further, qualitative analysts consistently expressed the need to retain agency and control rather than have models dictate and potentially bias their analyses by focusing on details and correlations irrelevant to their on-going exploration. Hence it is essential to provide the analyst with the ability to critically inspect the model and its recommendations.

In order to facilitate collaboration and model inspection, the analyst is presented with interactive visualizations to explore the entire dataset and to provide the user with explanatory analysis for debugging the system. We again model our visualizations to mimic an inductive (bottom-up) workflow by following the different levels starting from detail to overview similar to how manual qualitative analysts interact with data by starting from inspection of subsets of data and then constructing themes from them inductively. Our tool primarily focuses on presenting the analyst with explanatory visualizations explaining how and why the model has made each prediction. The interactive visualizations provide model transparency by first showing instance-level detail and moving towards a more corpus-wide visualization (c.f., Fig. 1).

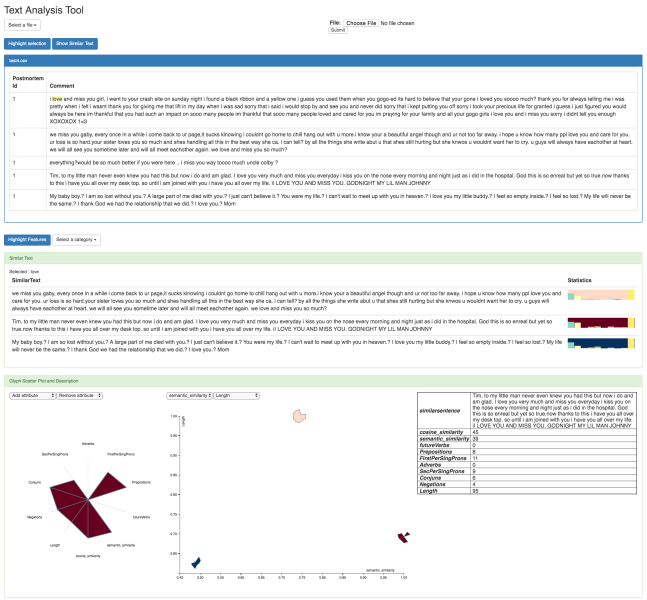
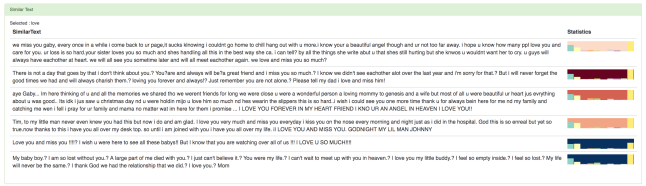


Figure 1: The preliminary QualVis interface consists of three separate components ranging from details about the current exemplar (full text, top) to suggested next exemplars (text sentences with sparkbars, middle) to corpus-scale feature comparisons (glyph-based scatterplots, bottom).

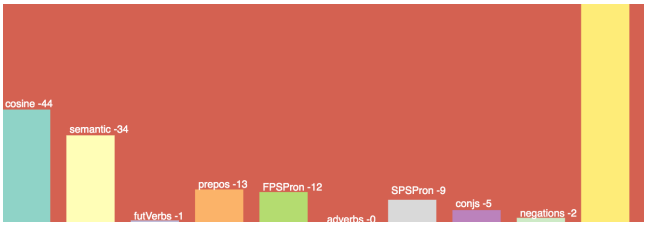
4 System Design

QualVis (Fig. 1) is a web-based tool built using Django, SciKitLearn, and D3.js. The visual analytics interface consists of a raw text display for coding relevant exemplars, a suggestions display containing recommended relevant exemplars and explanatory visualizations, and a corpus-scale view containing feature-based data about a large collection of exemplars. As in a traditional paper-based analysis, an analyst first highlights and codes a specific word, phrase, or sentence in the raw text view. The tool then classifies each instance as an emotional distress or not and computes similar occurrences to the highlighted text within the dataset based on the current model of the on-going analysis. The similarity prediction component is based on cosine similarity (on multinomial vectors weighted by TFIDF) and semantic similarity (using WordNet Sentence Similarity based on Semantic Nets and Corpus Statistics [12]).

In order to classify the similar instances, a Naïve Bayes model is trained on labeled social media data from 1000 MySpace comments with labels indicating whether each instance is an emotional distress or not from a prior inductive analysis of the dataset [2]. We use Naïve Bayes for two primary reasons. First, we wanted to create a simple to understand model to see if we can derive useful patterns using QualVis. Secondly, Naïve Bayes was more open to user-defined modifications through feature amplification than it's complex coun-



(a) The suggestions display



(b) Sparkbars

Figure 2: (a) The suggestions display provides the raw text of sentences from the corpus most similar to the most recently coded passage. Similarity is computed as a weighted sum of the label classification confidence and the cosine similarity of the text itself. (b) Sparkbars encode the normalized weight distribution of features predicting similarity, with the background color summarizing the overall classification confidence for a given recommendation.

terparts. This is especially beneficial for users who may be less familiar with the mathematics behind sophisticated classification algorithms. This model does not use the bag of words approach but rather uses a set of features as shown in Table 1. We use these features as they represent elements of the text that analysts can readily reason over (e.g., pronoun usage) as opposed to complex compound features or topics that lack a human-interpretable foundation. The features used here were selected based on manual analyses [2] as relevant for the corpus and confirmed during our development process using automated feature selection methods. Fig. 2 shows the results obtained from the tool. All the instances similar to a highlighted word "love" are identified by the classifier and displayed with inline sparkbars adjacent to them to communicate feature weightings to the analyst.

4.1 Document Level Visualization

After the system computes similar passages in the text, these passages are visualized alongside explanatory visualizations intended to communicate aspects of why a particular passage was recommended (Fig. 2). The spark bars represent the normalized score of all of the features as described in Table 1. The background color of the spark bars indicate the class and confidence of a given suggestion ("distress" or "not distress") computed using the classifier. These scores allow the analyst to reason about why the passage was recommended, allowing the analyst more agency in choosing whether or not to follow the model's recommendation.

A diverging color scale from ColorBrewer[4] is used to code the classifier confidence and classification label. A darker shade of blue indicates that the classifier is more confident the document is "not an emotional distress". Similarly, a darker shade of red indicates that the classifier is more confident the document is "an emotional distress".

We use the background of the sparkbar to provide rapid insight into how strongly a recommended passage does or does not reflect the most recently applied code as analysts can quickly assess whether an example provides a positive exemplar or counterfactual. The lightness of the background enables a collaborative assessment of the example by surfacing the model's confidence in its recommendation. Analysts can use this information to focus on finding more elements either as added exemplars or potential counterfactuals to an on-going analysis. Analysts can use the sparkbar representation to more critically inspect a given instance and why the automated system may find the passage of interest given the most recent coding in the raw text. These recommendations allow the analyst to make a holistic and informed decision as to how they wish to navigate the text corpus, reflecting on the features and potential biases that led to the given recommendation before following the algorithm's guidance.

[4] http://colorbrewer2.org

4.2 Interactive Corpus Scale Visualization

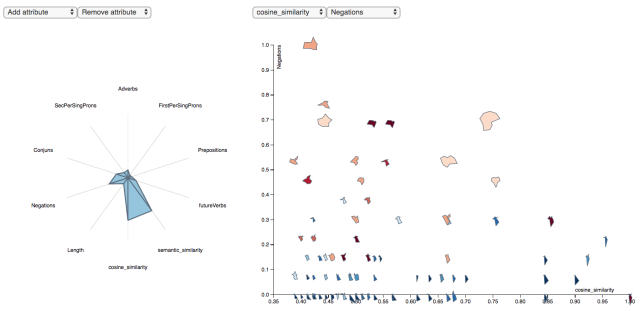
While our specific recommendations help to identify relevant additional exemplars from the corpus, analysts may instead wish to use a more traditional overview approach to supplement their exploration. Given that the features used in the classifier for a given code reflect the algorithm's learned impressions of the analyst's current linked passages, we anticipate the corpus scale visualizations will best support inductive methods by focusing on patterns across these features. In QualVis, analysts can look at the distribution of key features used in the classification for a given code using a glyph scatterplot.

Each glyph represents the complete feature set of every document in the dataset. These star glyphs have each of the 10 features as their rays with the length of the ray defining the intensity of that feature in a document. These glyphs are then plotted on a scatter plot whose axes can be set to one of the 10 features (Table 1). Hovering over a given glyph loads into a detail view which shows the shape of the glyph at larger size mapped to labeled axes and provides additional details about the document in a text table. The glyphs are color coded similar to the background of the sparkbars. We use star glyphs to allow for multidimensional feature representation (the axes of each individual glyph) while retaining a one-to-one mapping between glyphs and documents.

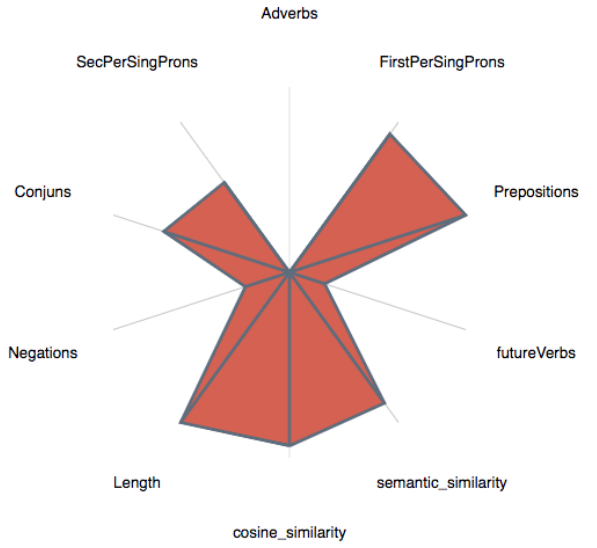
Placing these glyphs in a scatterplot allows analysts to quickly cluster documents on the most critical features (mapped to the x- and y-axes) while still comparing the relative feature distributions by attending to the general shape formed by each document in the corpus. For example, in Figure 3, the orientation of blue colored glyphs in the lower right of the scatterplot suggests documents with similar feature distributions with respect to indicators of emotional distress. Conversely, in the same region, we find two red colored data points with very similar shape to the blue colored data points surrounding them. This indicates outliers that may contain interesting counterfactuals to the most recently labeled code that may offer interesting passages for the on-going exploration. The goal of this glyph based scatterplot is to allow the users to analyze the interplay of features within a document and also allow them to compare the same between documents through one single visualization. For this reason, we first present the analyst with all the features mapped to the glyphs and all the glyphs mapped to the scatterplot from which the user can deconstruct to a smaller subset of features if necessary.

4.2.1 Interactive Machine Learning

An analyst can also influence the model by highlighting words, sentences, or phrases representing a given qualitative category from the similarly identified text and assigning that passage a given code using a dropdown list (Fig. 5). The model can then retrain with the newly provided code to update its recommendations based on the most recent state of an on-going analysis. This coding process mimics that used in common tools like MaxQDA by allowing analysts to interactively highlight and categorize specific passages of text. However, the approach goes one step further by adding the newly coded data to the model as a labeled example of the specified code. By adding this data, the model can use active learning to update its



(a)



(b)

Figure 3: (a) A glyph scatterplot encodes corpus scale patterns within the dataset, allowing analysts to readily recognize patterns across relevant features by examining the shape or position of different star plot glyphs (b) Detail view of glyph shape.

understanding of the analyst's conception of the provided code in an effort to provide more relevant guidance towards related exemplars. We additionally employ feature amplification during highlighting where we increase the weighting of a given feature if it is contained within a highlighted passage. This weighting allows us to more actively consider specific information provided by highlights in partial passages (e.g., an interesting word within a broader sentence).

Most active learning approaches make use of an oracle, to aid with labeling tasks. The user is presented with data and applies a label to that data in an effort to improve classification performance. While this approach is useful in traditional machine learning, qualitative inductive analysis is not driven by "right" or "wrong" labels: the concept of what is represented by a code often emerges as a result of the labeling process. An analyst often is not looking to simply find and predict examples of a code but rather to learn from and develop theories characterized by exemplars. By decomposing these codes into their component features, our visualizations offer analysts a chance to reflect more critically on aspects of text that are (or are not) truly relevant to a given classification and on their own intuitions as captured by these codes.

To that end, QualVis allows analysts to interactively add and remove features from the model to provide a feature-forward method

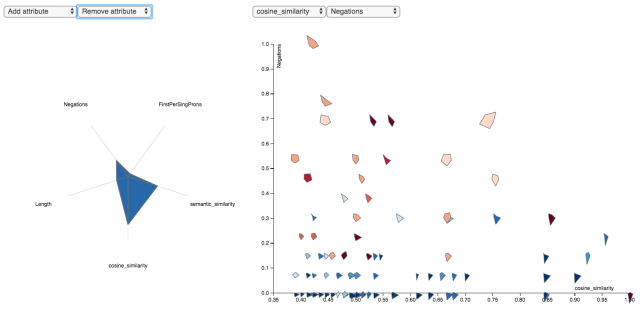


Figure 4: Glyph scatterplot from Fig. 2 with five features removed. By removing a particular attribute, it is possible to obtain two similar looking glyphs which were originally dissimilar in Fig. 2. Through this, we can infer the interplay of features in our model.

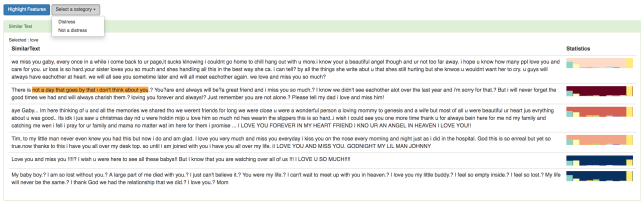


Figure 5: A view of the similarly identified text as identified by the model. An analyst can highlight interesting features and assign it a code using the given dropdown. Coded passages are fed back to the learning algorithm using Interactive labeling and feature amplification techniques.

for model refinement. This is done with the glyph structure in the scatterplot visualization. The users can add or remove any feature from the glyph which will change the shape of the glyphs plotted over the scatterplot and update the corresponding model.

5 Preliminary Use Cases

We are in the process of conducting a formal design review of QualVis with qualitative analysts to understand how well the current system features and design support QIMs. We hope that the outcomes of these discussions will illuminate more concrete consideration of how to integrate guidance and transparency through visual analytics systems in ways that preserve the agency and expertise core to QIMs. In preliminary explorations for these discussions, we found two concrete use cases that illustrate patterns in emotional distress generated using our approach.

Case 1: The usefulness of the glyphs by themselves as well as when they are coupled with a scatterplot is evident from the results obtained using the tool. Fig. 3 and Fig. 4 show the distinction between the two classes of prediction: documents cluster on the presence or absence of emotional distress both from the position along the feature axes and with respect to the broader set of features indicated by the glyph shapes. For example, we can see how all the glyphs in one class (colored blue) have a more or less similar shape. Coupling star glyphs with feature-based axes allows analysts to explore whether the patterns arising from their coding practices are well-captured by specific text features or if they represent more holistic combinations of features. Further, they can use these patterns to guide their explorations towards passages that either characterize the current qualitative theme (e.g., those passages indicative of emotional distress) or standing as outliers to the current theme (e.g., passages sharing a comparable feature distribution, but differing along a key feature or in classification strength).

Case 2: Fig. 6 shows how the confidence of the classifier varies

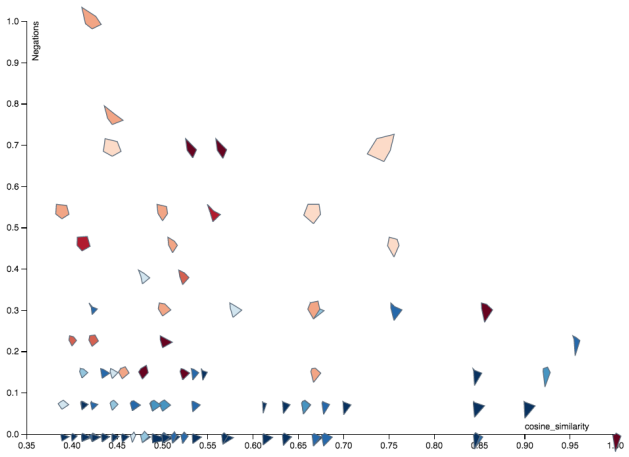


Figure 6: By mapping classification confidence (color) against the number of negations found in a passage (y-position), we find that the two features are inversely correlated: the model generated from the qualitatively coded data is less confident with data containing high numbers of text negations.

as a function of a specific feature (the number of negations in the passage). By plotting negation count on the Y-axis, we can see a correlation between the confidence of classifier and number of negations. The higher the number of negations, lower is the confidence of the classifier. This correlation is indicative of an interesting pattern in the underlying analyzed text: passages with a high number of negations lead to more confidence that a particular passage indicates emotional distress. Analysts can use this correlation to reflect on whether this pattern makes sense in the context of emotional expression online or if it might represent a bias present in the analyst's generative coding process.

Case 3: Fig. 1 is a full view of the QualVis interface which shows the documents that analysts can highlight followed by raw text suggested by the tool with inline sparkbars adjacent to them and the glyph based scatter plot at the bottom. In Fig. 3 we saw a cluster of similarly shaped glyphs on the bottom right corner of the scatterplot with only one of them belonging to a class that is different from its neighboring ones. If the analyst hovers over that particular glyph, they should see a pop up of the text of the document associated with that glyph. From this view they can manually corroborate the model's classification of this document.

In this example, the classification made by the model is either correct or incorrect. In the event that the classification is appropriate, the presence of such an outlier could imply that not all the features used in the model hold significant value to this particular document since all the other similarly shaped glyphs belong to the opposite class. To emphasize the features that strongly identify this document to its class grounded in an analyst's intuitions, the user can highlight the relevant words/sentences in the raw text view. On the contrary, if the model's prediction is identified as inaccurate by the analyst, they can go to the raw text view and highlight the words/sentences in that document that the analyst deems fit and select the appropriate category which will then emphasize the feature values in the model. This emphasis also allows the model to suggest new passages in the document-level visualization to further refine the on-going model and enhance the overall analysis.

In this use case, an analysts' interaction with the tool through initial highlighting of interesting words/sentences, feedback through the emphasis of relevance of words/sentences to a particular class and addition or removal of features in the glyph view allows an analyst to collaborate with the algorithm to refine their own thinking and to improve the overall quality of the analysis without sacrificing analyst agency.

6 Open Challenges

In this paper, we explore how visualization systems can mine analyst annotations to scale up qualitative inductive analysis in the context of emotional distress. We use Naïve Bayes to let analysts explore text corpora for sparse phenomena by coupling inherent statistical features of text and provide preliminary explanatory visualizations that allow analysts to interpret algorithmic recommendations and uncertainties through oracle-initiated feedback.

The work discussed here represents only preliminary steps towards understanding how visual analytics approaches can support qualitative analyses at scale. For example, most qualitative analyses are not likely to be as simple as a binary classification task. The simplest way to handle the multi-class classification problem for large datasets is to model a set of binary classifiers where a data point either belongs to a class or it does not. While our approach allows for analysts to consider multiple codes, the suggestion views can only use whether or not one particular code applies to a passage. One potential problem with this method is the assumption that the labels are not mutually exclusive and that some data points may belong to multiple classes whereas some others may not be relevant to any label. While the latter scenario can be useful when sifting through huge corpora, the relevance of the former scenario heavily relies on the analysts' requirements. Read et al. [15] uses a novel classifier chains method that can model label correlations while maintaining acceptable computational complexity that may offer preliminary support towards a true multiclass guidance system.

Further, a details-first visualization approach can lead to *desert fog*, a phenomenon where an analyst can become lost with respect to their current workflow as they move across different levels of detail [7]. Given that our system focuses on user guided data exploration, it is important to present analysts with enough information to navigate through the corpus while not losing sight of relevant information necessary for theory building. Avoiding desert fog in inductive visualizations is a critical aspect of future work: while methods like critical zones [7] may help orient analysts, visualizations should exercise caution when anchoring analysts to avoid biasing workflows through algorithmically reduced global overviews removed from the semantics of the actual exploration.

The strength of our approach lies in helping analysts navigate large corpora at different levels of detail and monitoring the links and codes applied to the data in real-time to focus analysts' energies according to inductive processes. We augment these recommendations using basic explanatory visualizations to support critical reasoning around the classification, but such explanatory influences may inadvertently cause analysts to put too much trust in the system [13]. This could in turn gravitate an analysts' interests to align with the model's discovery thereby corrupting our collaborative feedback mechanism. Future work is needed to understand how these systems can maximize information gain for the analysts while minimizing the bias generated from transparent guidance.

By addressing the issues raised in this section along with further evaluation, we envision the system to be useful in domains that traditionally lack the application of conventional computational analysis methods such as public health [14], law and policy, and disaster response. Currently, the choice of features in our model is specific to the task of emotional analysis in social media data. Our team is working on techniques for generalizing and automating the feature selection process. We are working with collaborators across these domains to identify potential use cases for generalizing our approach. The visualization methods created will need to intuitively communicate computational models and uncertainty to inform reasoned decision making around large scale data, allowing people to refine and apply the outcomes of statistical products even without statisti-

cal expertise. We are currently in the process of using QualVis to better understand how visual analytics might support scalable QIMs and for designing extended computational and visualization methods for expert-driven inductive analyses.

Acknowledgments

This research was supported by NSF Award #1764089. We thank Jed Brubaker for his support of this work.

References

[1] J. R. Brubaker and G. R. Hayes. "we will never forget you [online]": an empirical investigation of post-mortem myspace comments. In *Proc. CSCW*, pp. 123–132. ACM, New York, 2011. doi: 10.1145/1958824.1958843

[2] J. R. Brubaker, F. Kivran-Swaine, L. Taber, and G. R. Hayes. Grief-stricken in a crowd: The language of bereavement and distress in social media. In *Proc. ICWSM*. AAAI, 2012.

[3] S. Chandrasegaran, S. K. Badam, L. Kisselburgh, K. Ramani, and N. Elmqvist. Integrating visual analytics support for grounded theory practice in qualitative text analysis. In *Computer Graphics Forum*, vol. 36, pp. 201–212. Wiley Online Library, 2017.

[4] M. Choi, C. Park, S. Yang, Y. Kim, J. Choo, and S. R. Hong. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proc. CHI*, p. 230. ACM, 2019.

[5] C. Felix, A. Dasgupta, and E. Bertini. The exploratory labeling assistant: Mixed-initiative label curation with large document collections. In *Proc. UIST*, pp. 153–164. ACM, 2018.

[6] P. Goffin, J. Boy, W. Willett, and P. Isenberg. An exploratory study of word-scale graphics in data-rich text documents. *IEEE Transactions on Visualization and Computer Graphics*, 23(10):2275–2287, Oct. 2017. doi: 10.1109/TVCG.2016.2618797

[7] S. Jul and G. W. Furnas. Critical zones in desert fog: aids to multiscale navigation. In *Proc. UIST*, pp. 97–106. ACM, New York, 1998. doi: 10.1145/288392.288578

[8] A. Kangasrääsiö, Y. Chen, D. Głowacka, and S. Kaski. Interactive modeling of concept drift and errors in relevance feedback. In *Proc. UMAP*, pp. 185–193. ACM, New York, 2016. doi: 10.1145/2930238.2930243

[9] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proc. CHI*, pp. 5686–5697. ACM, New York, 2016. doi: 10.1145/2858036.2858529

[10] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proc. IUI*, pp. 126–137. ACM, New York, 2015. doi: 10.1145/2678025.2701399

[11] T. Luciani, A. Burks, C. Sugiyama, J. Komperda, and G. E. Marai. Details-first, show context, overview last: supporting exploration of viscous fingers in large-scale ensemble simulations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):1–11, 2018.

[12] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[13] K. F. Oduor and E. N. Wiebe. The effects of automated decision algorithm modality and transparency on reported trust and task performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 52, pp. 302–306. SAGE Publications Sage CA: Los Angeles, CA, 2008.

[14] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Proc. ICWSM*. AAAI, 2011.

[15] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333, 2011.

[16] B. Settles. *Active Learning*. Morgan & Claypool Publishers, United States, 1st ed., 2012.

[17] A. L. Strauss and J. M. Corbin. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, Inc., Thousand Oaks, California, 2nd ed., 1998.

[18] D. R. Thomas. A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2):237–246, 2006.