Automating Large-Scale Annotation for Analysis of Social Media Content

Katie Kuksenok, Michael Brooks, John J. Robinson, Daniel Perry, Megan K. Torkildson, Cecilia Aragon
University of Washington
{kuksenok, mjbrooks, soco, dbperry, mtorkild, aragon}@uw.edu

ABSTRACT
Analytic annotation can shed light on social phenomena embedded in digitally-mediated communication, but methods of manual annotation do not scale well to larger data sets and are not well supported by existing manual labeling tools. We present a set of visual text analytics tasks for maintaining transparency as large text datasets undergo analysis through manual and automatic annotation. The discussion is grounded in our ongoing research on analyzing expressions of affect in chat logs from a distributed scientific collaboration.

Author Keywords
Qualitative methods; visual analytics; computer-mediated communication

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces.

General Terms
Human Factors; Design; Measurement.

INTRODUCTION

Social media is increasingly integrated into organizational processes within corporations, non-profits, and government agencies [6], generating rich data for research on social phenomena. The data can be analyzed using qualitative methods, traditionally applied to materials such as interviews and field notes; however, methods requiring extensive human attention and interpretation do not scale robustly to large social media data sets.

Text analysis relies on manual annotation, or coding, whether as a precursor to quantitative analysis and hypothesis testing, or as part of the construction of grounded theories of social phenomena. Manual coding of large data sets is impractical; natural language processing (NLP) and machine learning (ML) algorithms may be used to automate annotation over the entire corpus based on a small set of manually labeled data. However, a methodologically sound implementation of such a process requires researchers to maintain a complete understanding of the dataset as it grows and changes. Visual analytics can support many relevant tasks involved: finding interesting subsets of data, understanding temporal relationships, and evaluating discrepancies.

Drawing on our research into affect in distributed scientific collaboration chat logs, we outline visual text analytic tasks related to annotating social media datasets, understanding these annotations as they evolve over time, scaling them through automated classification, and analyzing the results. Although examples focus on chat logs, the discussion can be extended to many types of social media content.

AFFECT IN DISTRIBUTED COLLABORATION

Our research focuses on the expression of affect in text chat used by distributed teams of scientists. We use *affect* to refer to an inclusive concept that spans emotions and feelings distinct from cognition [11], more pervasive than the neurophysiological experiences of emotions [8]. Expressions of affect and emotion have been linked to cooperation, performance and creativity within a work environment [1,5]. We aim to develop a better understanding of the role of affect in team dynamics.

Our chat dataset was collected from the Nearby Supernova Factory, an international astrophysics collaboration, over four years [2]. Out of 485,045 total chat messages, the top 32 human participants contributed over 500 messages each, or 300,684 messages total. Individual chat messages are short, often between 5 and 10 words in length. The scientists frequently use jargon in addition to unusual grammar and spelling. Topics of conversation vary, including technical conversations about equipment, discussion of scientific results, and socializing.

We labeled each message, combining open coding and relevant literature in a modified grounded theory approach [12]. It took eight weeks, three main coders, and five additional coders to label 5% of the data, illustrating the impracticality of manually coding the entire dataset. We have begun developing automated methods for identifying affect expression based on manual annotations. Statistical classification methods can detect the overall positive or negative sentiment of long, relatively well-formatted blogs, articles, and online text [4,7,13]. Similarly, recent work has classified text from social network sites, blogs, and discussion forums [10,14]. There are challenges specific to informal text but many of these approaches are promising.

Datasets produced by social media interactions are detailed and span a great range of social phenomena. Machines can hardly be expected to achieve the level of understanding and interpretation encoded in manual labels [10]. However, the availability of tools such as machine learning marks a methodological shift that deserves care and discussion, but also carries tremendous potential for new understanding of

large data sets. In our discussion of visual text analytics tasks, we stress the importance of transparency.

TASKS

The following tasks concern many stages of analysis, from open coding to analysis of hybrid manually/automatically labeled datasets. Each involves scaling existing practice to the volumes of text data generated by social media.

Collaborative coding of structured data

The annotation process can be augmented with visual presentation of structure and context of the data being coded. Unlike interview transcripts and field notes, social media produces more heterogeneous text traces. Conversations include many recurring individuals over time, and may include additional metadata, such as locations or threading. The annotation task becomes more complex, with the need to see how a particular snippet of content relates to all the other snippets in time, authorship, and any available information of conversation hierarchy.

We developed a tool (Figure 1) for coding large amounts of chat data more efficiently. Visualizing message frequency over time characterizes the pace of conversations. Color-coding speakers makes the conversation thread available at a glance. The color-coded grid of codes helps users remain cognizant of the range of available affective categories. The colors are based on Plutchik's taxonomy of emotion [9].

Visualization can also help support collaboration. Teams of coders need to understand how each coder is annotating the data. Visual representations of statistics like inter-coder agreement would help researchers quickly identify problem areas in the coding scheme or the coding process. Linking these visualizations to selected examples of coded chat messages from the data set would allow researchers to better understand the nature of disagreement, guiding them towards more effective solutions.

Understanding structures of open coding

Qualitative analysis progresses from open coding towards structured, axial coding. In open coding, coders read data and create new codes as needed, allowing important themes to emerge organically, rather than imposing a predefined theory on the data [3]. Coding is accompanied by extensive discussion and reflection, and as more data is coded, the coding system gradually becomes more structured.

The basic mechanics of this process are hindered by the size of social media datasets. We found that effective critical reflection and consensus building is challenging when a large number of codes have been applied in an open-coding process across different sections of a large, diverse data set. Visual text analytics tools can help make the meanings and relationships between the codes more transparent, facilitating effective discussion.

We have begun to explore tools to support this process, including a tool for retrieving a subset of messages where a selected code has been applied, displaying each result in the

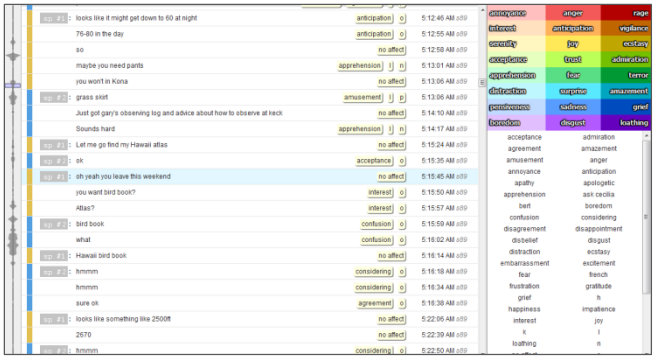

Figure 1: Visual tool supporting collaborative coding for affect in chat messages.

context of surrounding conversation. In one instance, we observed that very few messages had been coded as *serenity*; from discussion it was clear that its meaning was too enigmatic, and coders' definitions for *serenity* ranged from "pleased" to "peaceful" and overlapped with another code, *acceptance*. Using this tool, we could identify messages where *serenity* had been applied and reach a shared understanding of its meaning, leading to more frequent and appropriate uses of the code. A shared understanding of the open coding process among the researchers is crucial as the coding scheme stabilizes.

In addition to finding subtle differences between coders, visual text analytics tools could address related challenges, such as identifying and understanding code overlap – which codes seem to have similar meanings based on how they are applied. Understanding how the meanings of codes change over time as coding becomes more structured is another important task. Visualization exposing these shifts would provide insights into the process, and could allow the user to drill down to text examples for further interpretations and actions. In some cases, discovering the associations between codes can even be the sole analytic objective.

Understanding changes to code system

Open coding involves changing the coding scheme as coding progresses. Even as the coding scheme gains structure and stability, changes can still occur when new phenomena are encountered in the data. With multiple coders making changes, two codes with similar meanings may be created in different contexts (duplication), or a single code may be used in different ways (ambiguity). In the case of duplication, researchers may merge the two similar codes into one. Or, if a code is ambiguous, it may be desirable to tease it apart into multiple codes. But what effects will such corrections to the coding scheme have?

Even if two codes are near duplicates, merging them may create inconsistency if they are used with slightly different connotations. For example, merging a code from Plutchik's emotion taxonomy for *joy* with our original *happiness* code was difficult because definitions overlapped. Some felt that *joy* was more enthusiastic, but others did not agree. Such inconsistencies must be understood before two codes can be

merged. Visual text analytics can provide researchers with means to explore the effects of a change to the coding system prior to enacting it, through visualizations of code coincidence, code distribution over the data in terms of time and chat participant, and cluster analysis providing the ability to drill down to specific examples representing different senses in which codes have been used.

Developing useful features

To enable automatic annotation, we must extract useful numeric features from the text. Many choices are available, such as what words to count; how to count punctuation, whether to collapse synonyms; whether to analyze a paragraph, a sentence, or a sliding window of messages. These choices are pivotal to the effectiveness of ML.

The process of developing a useful feature set begins with hypothesizing high-level features, best informed by domain familiarity. For example, after reading a large amount of chat data during coding, we hypothesized that punctuation might be an effective predictor of affect expression. Once a high-level feature has been identified, it must be operationalized (e.g. the presence or absence of "???" or "!!!"), requiring iterative evaluation and adjustment to ensure the feature is extracted as intended. Every feature must be evaluated, as each feature added to the set increases the risk of over-fitting and the training time for classifiers.

Visual analytics tools can support the rapid evaluation of new features. In considering each feature, we estimated its usefulness based on its distribution over our data relative to different codes, as well as changes over time, differences between chat participants, and correlations with other features. To support rapid iteration and development of a quality feature set, visual representations of these measures, linked to the underlying text, would be invaluable.

Locate interesting areas of the text stream

At various stages in our analysis, we sought new areas of the dataset with specific desirable properties. With large datasets, it is typical to limit manual coding to a subset of the dataset. A random subset is limiting for more qualitative analyses because important but rare phenomena may be missed, so a carefully targeted approach can be preferable.

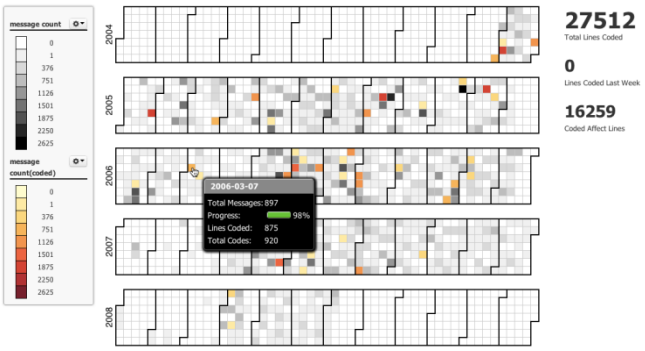

Figure 2: A calendar view of the dataset.

We visualized message density over time (Figure 2) to identify subsets that matched our criteria, aiming to evenly distribute our effort over the 4.5 year data set. Within each year, we attempted to find sections that were primarily in English, appeared to be rich in certain forms of affect, and were not too sparse. Selection of new passages was time-consuming and methodologically unsatisfying. Visual tools that take advantage of the feature set and existing manual labels would be helpful for the task of choosing what sections of the dataset to annotate next.

Understand temporal relationships

Analyzing sequences of codes over time reveals patterns in common behaviors, or how real-world context affects the manifestations of behaviors. This is the primary objective in our research. Visual tools can support understanding of temporal relationships in a large annotated dataset.

Codes may recur in a temporal pattern. One occurring shortly before or after another may indicate properties of the section of text that would be valuable to identify. In our dataset, *annoyance* may precede a series of *anger* or *disgust* code instances as a disagreement escalates. These temporal patterns characterize social processes, and visualizations showing the flow of related codes can help researchers

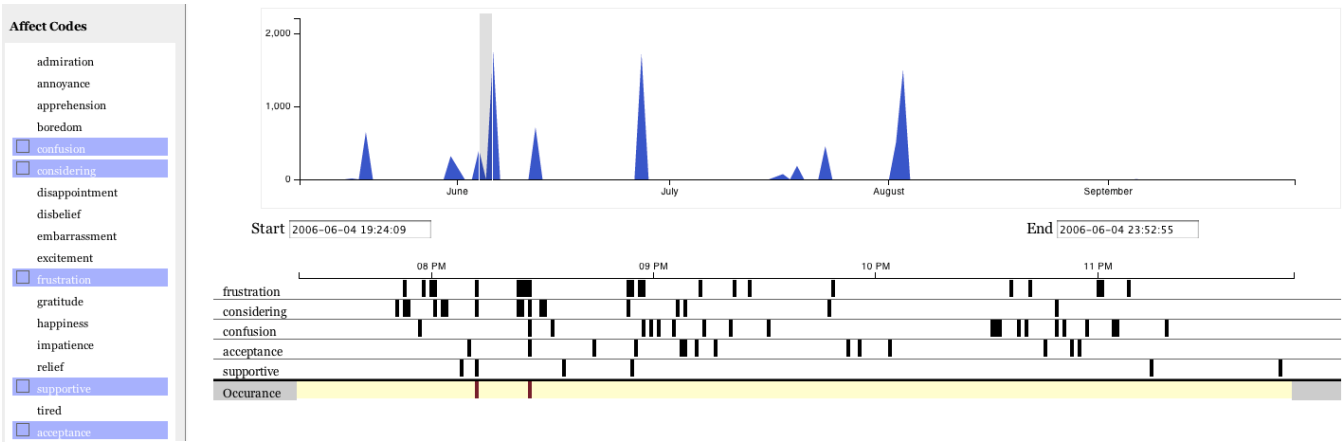

Figure 3: An interactive visualization for exploring temporal relationships between codes. Each line shows individual code occurrences; the bottom line shows occurrences of the selected codes to quickly identify correlations.

discover and understand social phenomena. We visualized affect code occurrences along an adjustable timeline (Figure 3) to identify temporal relationships between many codes at once, as global trends and at finer granularities.

Evaluate discrepancies

In manual coding, an iterative process of developing a coding scheme and training coders helps establish consistency, validated by metrics for inter-coder agreement. At a large scale, with both manual and automatic annotation, knowing discrepancies between automatic and/or manual labelers only in terms of a single statistic provides little useful information, failing to effectively identify and understand inconsistencies. Visualizing inter-coder reliability metrics, alongside code occurrences and distributions of descriptive features, can help users spot patterns, such as disputed codes or features of difficult-to-interpret text. The causes of these inconsistencies could be verified through on-demand access to the raw text. Descriptive analysis of reliability at scale is vital to annotation as an iterative process, characterized by improving coding for greater reliability and analytic power.

CONCLUSIONS

We have defined a set of visual text analytics tasks for researchers working with large text datasets produced by social media. These tasks arose from our own work conducting open coding and automated annotation for the analysis of affect in chat data produced by distributed scientific collaboration. We highlight transparency and maintaining provenance across transformations, as well as making raw data available on demand. These and other, task-specific considerations are motivated by the need to make automation effectively scale manual annotation as an analytic tool for large volumes of social media data, as well as by interest in maintaining methodological validity throughout this process.

ACKNOWLEDGMENTS

We thank the scientists of the SNfactory collaboration. This work was funded in part by an NSF Graduate Research Fellowship in Computer Science.

REFERENCES

1. Amabile, T.M., Barsade, S.G., Mueller, J.S., and Staw, B.M. Affect and Creativity at Work. *Administrative Science Quarterly 50*, 3 (2005), 367–403.
2. Aragon, C., Poon, S., Monroy-Hernandez, A., and Aragon, D. A Tale of Two Online Communities: Fostering Collaboration and Creativity in Scientists and Children. *Proc. C&C 2009*, ACM (2009), 9–18.
3. Charmaz, K. *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE, London, 2006.
4. Gill, A.J., French, R.M., Gergle, D., and Oberlander, J. The language of emotion in short blog texts. *Proc. CSCW 2008*, (2008), 299–302.
5. Grandey, A. Emotions at Work: A Review and Research Agenda. In *Handbook of Organizational Behavior*. SAGE, London, 2008.
6. Jue, A.L., Marr, J.A., and Kassotakis, M.E. *Social Media at Work: How Networking Tools Propel Organizational Performance*. Jossey-Bass, San Francisco, CA, 2010.
7. Keshtkar, F. and Inkpen, D. Using sentiment orientation features for mood classification in blogs. *Proc. NLPKE 2009*, IEEE (2009), 1–6.
8. Moore, B.S. and Isen, A.M. *Affect and Social Behavior*. Cambridge University Press, 1990.
9. Plutchik, R. The Nature of Emotions. *American Scientist 89*, 4 (2001), 344–350.
10. Rosé, C., Wang, Y.-C., Cui, Y., et al. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *Computer-Supported Collaborative Learning 3*, 3 (2008), 237–271.
11. Russ, S.W. *Affect and Creativity: The Role of Affect and Play in the Creative Process*. Routledge, 1993.
12. Scott, T.J., Kuksenok, K., Perry, D., Brooks, M., Anicello, O., and Aragon, C.R. Adapting Grounded Theory to Construct a Taxonomy of Affect in Collaborative Online Chat. *Proc. SIGDOC 2012*, (2012).
13. Tausczik, Y.R. and Pennebaker, J.W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology 29*, 1 (2009), 24–54.
14. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology 61*, 12 (2010), 2544–2558.