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

Introduction

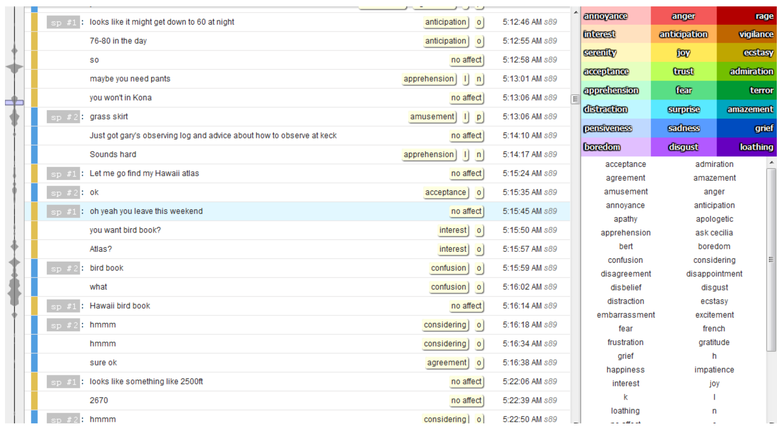
Qualitative methods do not scale well to the sheer volume of rich social media data, as it is time-consuming to apply human interpretation at this scale. Natural language processing (NLP) and machine learning (ML) can be used to automate coding based on a subset of manually labeled data. However, researchers must maintain understanding of the data set. Visual analytics can assist comprehending, detecting patterns in, and analyzing vast text data sets.

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.

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, more pervasive than the neurophysiological experiences of emotions. 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. In 485,045 total messages, there is frequently jargon in addition to unusual grammar and spelling. Topics of conversation span technical conversations about equipment, discussion of scientific results, and socializing. We labeled each message in 5% of the data, and have begun developing automated methods for identifying affect expression based on manual annotations.



We developed a tool for coding large amounts of chat data more efficiently. Visualizing message frequency over time shows 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; colors reflect the mapping in Plutchik's taxonomy of emotion.

Visual Text Analytics Tasks

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 stress transparency and maintaining provenance across transformations, as well as making raw data available on demand, in order to make automation effectively scale manual annotation as an analytic tool for large volumes of social media data.

Collaborative coding of structured data

Unlike interview transcripts and field notes, social media produces text traces with additional metadata that should be revealed during annotation. Our tool (left) expedites collaborative coding of chat logs.

Understanding structures of open coding

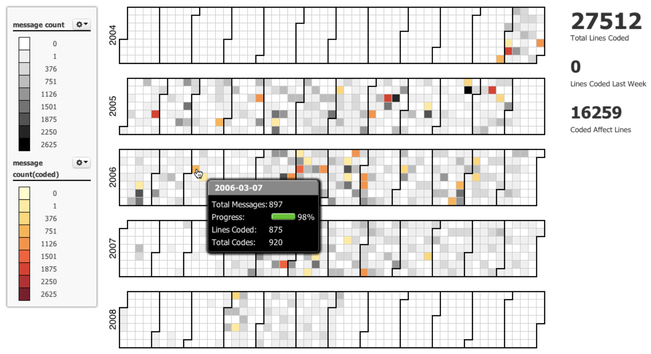
Qualitative coding is accompanied by extensive discussion and reflection, and as more data is coded, the coding system gradually becomes more structured. However, effective critical reflection and consensus building is challenging when a large number of codes have been applied in an open-coding process across different sections of a large, diverse data set.

Evaluating discrepancies

Visualizing inter-coder reliability metrics, alongside code occurrences and distributions of features, can help users spot patterns, such as disputed codes or features of difficult-to-interpret text.

Locating interesting areas of the text stream

With large datasets, manual coding is limited to a subset of the dataset. A random subset is limiting for qualitative analyses because important but rare phenomena may be missed, so a carefully targeted approach can be preferable. For example, we chose data to analyze based on message density during a day's log (below).

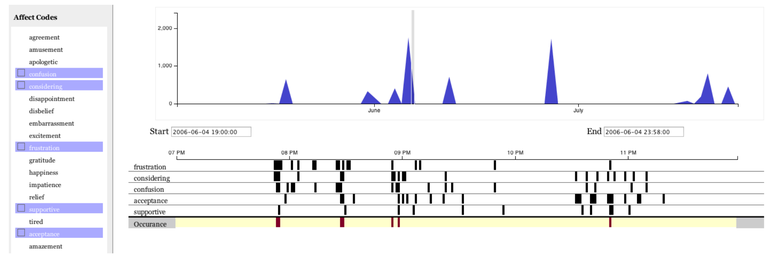


Understanding changes to code system

Changes to the coding scheme can 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). Finding these conflicts as well as enacting changes to resolve them requires exploring 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.

Understanding temporal relationships

Analyzing sequences of codes over time reveals patterns in common behaviors. Visual tools can support understanding of temporal relationships in a large annotated dataset.



Interactive visualization of temporal relationships between codes. Each line shows individual code occurrences; the bottom line shows occurrences of the selected codes to quickly identify correlations.

Developing useful features

Automation requires extracting useful numeric features from text. Many choices are available, and are pivotal to the effectiveness of ML. The rapid evaluation of a new feature involves examining its distribution over our data relative to different codes, as well as changes over time, differences between chat participants, and correlations with other features.

For more information, contact Katie Kuksenok at kuksenok@cs.uw.edu or Cecilia Aragon at aragon@uw.edu or visit http://depts.washington.edu/sccl

Scientific Collaboration & Creativity Lab

HCDE Human Centered Design & Engineering

dub

UNIVERSITY of WASHINGTON

University of Washington Computer Science & Engineering