# Human Centered Tools for Analyzing Online Social Data

Michael Brooks

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Dr. Cecilia R. Aragon, Chair

Dr. Mark P. Haselkorn

Dr. Sean Munson

Dr. Kate Starbird

Program Authorized to Offer Degree:

Human Centered Design and Engineering

University of Washington

## Abstract

Human Centered Tools for Analyzing Online Social Data

Michael Brooks

Chair of the Supervisory Committee:

Associate Professor Cecilia R. Aragon

Human Centered Design and Engineering

In the social sciences, researchers are increasingly turning to datasets collected from social media, online chat, forums, and email to address questions about human communication and behavior. However, these datasets are notoriously difficult to work with. Social media and online communication datasets push the limits of traditional research methods, force researchers to learn an array of new data science skills, and limit open and equitable participation in this important new research area. While this problem has many sides, one of the most significant challenges is a lack of technological support for online social datasets and mixed methods data analysis processes. Many researchers in this area have to create custom scripts and software for gathering, analyzing, and visualizing their data.

Solving this problem depends on understanding the data analysis processes and practices of social scientists working with online social data. In this dissertation, I present an ethnographic interview-based study on the work practices of researchers applying mixed methods to social media data, in order to better understand their data collection and analysis processes and generate implications

for design. Even with a good understanding of how social scientists work with data, significant questions remain about how to design helpful software. Based on a year-long engagement with a research group studying emotion in a large chat dataset, I discuss the implications of applying machine learning technology to "amplify" and scale up qualitative analysis from a small manually-coded set to the full corpus. Finally, I discuss two human-centered design projects focused on supporting aspects of the data analysis process: visual exploration of Twitter data, and collaborative qualitative coding of chat messages. This dissertation offers a descriptive understanding of how social scientists actually work with complex social media and online communication datasets, implications for designing better machine learning, visual analytics, and qualitative analysis software, and several open-source tools for analyzing online social data.

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgments

I would like to thank all of the participants who generously shared their time and knowledge, and without whom this dissertation would not be possible.

I am also grateful to my adviser, Dr. Cecilia Aragon, for her support and guidance, and to my committee for their insightful feedback. Thanks also to my internship supervisors who shared their valuable experience and mentorship.

Almost all of my research has been done in collaboration with others, so I owe a great debt to the many colleagues and friends with whom I've had the pleasure of working over the past few years. More specific acknowledgments of their contributions to the work presented in this dissertation are provided within each chapter.

Thanks also to my partner Tiffany, and to my family for their love and encouragement.

My research been supported by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation through the Data Science Environments program, and the National Institute of Standards and Technology.

# Chapter 1: Introduction

In recent years, social network sites and other online platforms have created the conditions for an explosion of research investigating online communication and behavior. Data about interactions between people, obtained from online sources such as Facebook, Twitter, or Wikipedia, are being used in business, journalism, government, and science, in what has been called a "social data revolution" (Weigend, 2009).

Within academia, the availability of online social data has opened the door for entirely new disciplines and attracted researchers from a variety of backgrounds. Changes are taking place in multiple academic communities, including communication (boyd, 2007), organizational systems (Diesner, Frantz, & Carley, 2005), information and computer science (Goggins, Laffey, & Gallagher, 2011; Starbird & Palen, 2012), and sociology (Zagheni, Garimella, Weber, & State, 2014). The work is diverse and difficult to summarize, but typically it involves the collection and analysis of records concerning online actions and interactions, often focused on a specific group, event, topic, or geographic location. Some researchers use this data to gain insight about offline behavior and social activity; others treat online behavior itself as the object of study; still others ask questions about the platforms, systems, and networks which support online activities. More detail about this body of work is provided in Chapter 2.

Just as there is great diversity in the approaches, goals, and backgrounds of social scientists working with online social data, there is great diversity in the data itself. Online social data might include anything from social media posts, to clicks on Facebook ads, to code changes pushed to GitHub. The data that are generated through users' interactions with online platforms and services differ in the type of platform where they reside or originate, the semantics of the action that

produces the data, the intentionality and goals of users in creating the data, and the types of content included (Manovich, 2011). In this dissertation, I focus on data analysis practices and design opportunities for social scientists working with a particular family of online social data: social media and online communication. This family of data is produced intentionally by users of online platforms and services with the goal of communicating with other people, and could include messages from Twitter, an organization's email history, a log of chat messages, or a hierarchical web of forum posts.

Despite the apparent differences among these data types, these media all include primarily text-based messages that encode rich, multifaceted information about users. Social media and online communication messages have an audience, whether small, large, or undefined, and can capture relationships between people and reveal social structures. Different kinds of messages often differ in their privacy expectations, message length, rate of delivery, and social context, but, while these differences are important, the commonalities between different message types mean that researchers studying them could be asking many of the same questions and employing similar methods and tools.

Although there is widespread interest in researching online interactions, access to social media and online communication data is limited in several ways. boyd and Crawford discuss a "digital divide" in the research community, where "those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access" (boyd & Crawford, 2012). Most private companies and organizations who collect and store online social data restrict access by outsiders, for economic, legal, ethical, and logistical reasons. Even if data is available, online social data is produced and delivered to researchers in great quantities, and some of the potential insights from these datasets depend on their size, breadth, and comprehensiveness. As a result,

manipulation and analysis of online social data requires computational work, which creates another kind of barrier: social data researchers often write scripts and software to obtain, explore, clean, filter, and analyze their data, which requires considerable time and knowledge.

Even with a command of data science techniques, analyzing online social data is a challenging task. These datasets are based on communication that occurs naturally in the world, so they vary unpredictably over time and space, and include unstructured informal text and other rich media that complicate analysis. The many facets of online social data require careful consideration and, often, require multiple analytical techniques and approaches. While many aspects of online communication data can be quantified and analyzed statistically or visually, the unstructured content in these datasets can also be considered qualitatively. In the social sciences, qualitative research and qualitative data analysis techniques provide researchers with powerful tools to study the nuances of human behavior in the real world (Denzin & Lincoln, 2011). Here, the researcher looks directly at the data, usually through a process of close reading, interpretation, and manual analysis, in order to answer complex questions about meaning, representation, practice, and other phenomena that are difficult or impossible to quantify.

In order to develop a more complete understanding of online social datasets, some researchers are beginning to explore mixed or hybrid methodologies that incorporate elements of both quantitative and qualitative traditions (boyd, Golder, & Lotan, 2010; Goggins et al., 2011; Maddock et al., 2015; Starbird & Palen, 2012; Vieweg, Hughes, Starbird, & Palen, 2010). By integrating close analysis of content and meaning with computational and statistical analyses, mixed methods allows researchers to understand their datasets both broadly and deeply (Creswell, 2014). However, mixed approaches to analyzing online social data present special challenges. In part, this is simply because mixed methods research requires a mastery of two distinct methodological approaches

(Johnson & Onwuegbuzie, 2004). Qualitative, quantitative, and computational methodologies involve different vocabularies, epistemologies, and goals. For those coming from backgrounds that do not traditionally include computer science training, learning scripting, statistical analysis, and other quantitative techniques can be a barrier. Meanwhile, for those with backgrounds in quantitative and computational fields, learning to see online social data from a qualitative research perspective is a challenge. Through various education and training initiatives, as well as interdisciplinary research collaborations, researchers in this new space are sharing skills and knowledge that support their use of mixed methods.

However, for online social data in particular, there are also significant technical barriers that inhibit qualitative analysis, and integration of qualitative analysis into mixed methods research. Qualitative research traditionally requires laborious manual inspection and interpretation of data (Coffey, Holbrook, & Atkinson, 1996), which is, practically, very difficult to carry out on large online social datasets. While a range of tools exist to assist researchers analyze traditional qualitative data (e.g. interviews, videos, and field notes), these qualitative data analysis (QDA) tools, like ATLAS.ti and NVivo, do not work well for social media and online communication data. Researchers taking a qualitative approach must work with the imperfect tools that are available, or build their own tools; both approaches create friction and impede research.

There are many other practical problems analyzing online social data where software might provide assistance. Exploration is a crucial part of data analysis, but the large size, text-based content, and extensive metadata in online social datasets makes effective exploration difficult. How can researchers get an overview of the data? How can they find interesting portions of the data that are relevant to their research questions, when the data are so diverse, dynamic, and noisy? Machine learning and other computational modeling techniques have the potential to bridge

qualitative and quantitative analysis in some cases (Janasik, Honkela, & Bruun, 2009; Rosé et al., 2008), but what sorts of machine learning should be used, and what considerations should researchers be thinking about? In some cases, collaboration can help by leveraging many eyes and diverse perspectives to improve analytical coverage (Heer & Agrawala, 2008), but how can groups of researchers effectively divide up a large social dataset, and how can they bring their interpretations together to produce shared understanding?

The adoption of Big Data approaches in the social sciences is having a transformative effect on many research areas, and the implications of these changes are broad: "[Big Data] reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality" (boyd & Crawford, 2012). In this time of transformation, understanding the needs and constraints for new technology to support emerging research practices can help ensure that our tools are enabling us to grow and explore new approaches. While there are many other factors at play, tools used for research can have a profound impact: "Researchers should pause and consider that technology is more than a tool [... it] requires researchers to reframe ideas about what can be done and how it is done [and] may have predetermined what is drawn to the researcher's attention" (St John & Johnson, 2000).

This dissertation is motivated by the question: how can data analysis tools be designed to enable more effective qualitative and mixed methods analysis of online communication and social media data? To better understand the context for technology design to support online social data research, I study the practices of researchers who are working with social media datasets, discuss a case study applying machine learning for qualitative online communication research, and consider the design of exploratory visual analytics tools and qualitative coding tools for online social datasets.

## 1.1 Methodology

This dissertation combines methods from human-centered design, ethnographic research, and machine learning. My overall goal is to contribute to the production of better tools and practices for social data research through understanding and solving practical design problems. Buchanan states that the designer's task is to explore "concrete integrations of knowledge that will combine theory with practice for new productive purposes" (Buchanan, 1992). In these terms, this dissertation involves designing integrations of computational techniques (e.g. machine learning and visualization) with qualitative and mixed methods practices. Drawing on Frayling's concept of "research through design" (Frayling, 1994), Zimmerman et al. have argued that *doing* design can contribute to the human-computer interaction community by framing under-constrained design problems, articulating a new desired state of the world, and producing models, prototypes, and products that serve as "design exemplars" and become a conduit for transfer between research and practice (Zimmerman, Forlizzi, & Evenson, 2007).

This dissertation takes a human-centered design approach. Human-centered design offers both a theoretical frame and practical tools for design that can maximize the usability, appropriateness, and experience of the designed solution for its intended users. Don Norman describes it as "the process of ensuring that people's needs are met, that the resulting product is understandable and usable, that it accomplishes the desired tasks, and that the experience of use is positive and enjoyable" (Norman, 2013). Methodologically, human-centered design implies a commitment to naturalistic observation of users in their native environments doing their normal activities, a kind of "applied ethnography," to inform the design process. As ideas and prototypes gradually converge towards a final product, a human-centered designer continually engages users in testing and iterative improvement.

In its commitment to naturalistic observation and user research, human-centered design draws on ethnographic research traditions. Many researchers in human-computer interaction and computer-supported cooperative work have used ethnographic and related field-based naturalistic methods both to inform the design of novel tools and to derive new concepts and theories about how people interact with technology (Lethbridge, Sim, & Singer, 2005; Luff, Hindmarsh, & Heath, 2000; Millen, 2000). In the field of visual analytics, concerned with the integration of data visualizations and computational analytics into interactive systems for analyzing data, researchers have called for more naturalistic studies to ground design and evaluations in real world problem domains (Munzner, 2009; Plaisant, Grinstein, & Scholtz, 2009). Ethnographic studies of situated activity in the workplace can reveal the processes by which work is accomplished, as well as the social, political, and material factors which structure these processes (Orlikowski, 2000).

Chapter 3 of this dissertation concerns an ethnographic interview study to understand the work of social data researchers practicing qualitative and mixed methods. In a traditional human-centered design process, such studies are conducted in the early phases of a design project and used formatively to guide problem selection and requirements (Norman, 2013). However, although it is presented first in this dissertation, this study was conducted *after* the systems discussed in Chapters 4, 5, and 6 had been designed, built, and evaluated. These systems, *ALOE*, Agave, and Text Prizm, were first created to meet the immediate needs of my research groups as we, ourselves, attempted to analyze social media and online communication datasets. After these projects were completed, I decided that a more systematic examination of work practices was needed to help illustrate the context in which these projects took place. Therefore, the design of the interview study is informed by my earlier experiences with *ALOE*, Agave, and Text Prizm, particularly in the selection of research questions and participants. It is presented first to give background material and concrete

demonstration of the need for better technology, and to organize and relate various aspects of social media data analysis into a coherent description that frames the subsequent chapters.

## 1.2 Research Questions and Contributions

Chapter 2 of this dissertation provides general background on research with social data in the social sciences, qualitative research, and computational methods. Chapters 3 through 6 explore the following three research questions:

- *What are the goals, barriers, tools, and processes associated with mixed methods online social data research?*
- *How can machine learning techniques be applicable for automation in mixed methods research with online communication data?*
- *How can tools be designed to help researchers explore and code large online communication and social media datasets?*

This dissertation makes three contributions:

- An improved understanding of the challenges and data analysis practices of social scientists using mixed methods approaches with social media datasets.
- Implications, considerations, and opportunities for designing and building data analysis tools for qualitative and mixed methods researchers working with social media and online communication data.
- A collection of open-sourced software tools, including *ALOE*, a machine learning tool for analyzing emotion in chat messages; Agave, a collaborative visual analytics tool for

Twitter data; and Text Prizm, a collaborative qualitative analysis tool for online communication data.

By improving our understanding of the practices of social scientists studying online social data, and developing implications and opportunities for design, this dissertation enables the creation of better tools and software for social scientists working with online communication data and points towards questions for future research. The tools created during this dissertation make it easier for researchers to qualitatively code large amounts of online communication data, to train and use machine learning to automate qualitative coding, and to visually explore social media datasets. These systems are available as open-source projects to promote further research and adoption by researchers. Better tools will make it easier for more researchers to explore the use of mixed methods and qualitative analysis of online social data.

# Chapter 2: Background

This dissertation is at the intersection of several disciplines, including social science research with social media and online communication data, qualitative research methods and practices, and computational analysis and machine learning. Therefore, in this chapter, I provide high-level background information to orient the reader unfamiliar with these areas. Below, I first review several examples of social science research that uses online social data. Next, I provide a brief history of qualitative methodology, and discuss its practical aspects. Finally, I explain several key concepts from machine learning, focusing on how classifiers are developed and evaluated.

## 2.1 The Social Data Revolution

Below, I will discuss examples of social science research that uses online communication and social media data. These few examples are not meant to be a thorough review, but to give a sense of the breadth and diversity of the field. The first group of examples studies the relationship between online and offline activities, while the second group approaches online communities and spaces as research sites in their own right.

### 2.1.1 Connecting Online and Offline

The general question of how activities conducted online might mirror, differ from, influence, and overlap with offline activities has been influential since the early days of the web. With the recent availability of online social data sources, researchers are asking we can learn about physical, face-to-face, and offline behaviors based on digital spaces, communities, and interactions. For example, public postings to social media sites could be used to infer population-level characteristics. Quercia et al. calculated a “gross community happiness metric” based on tweets originating from different

census communities in the UK, finding that their metric correlated with socio-economic status at the community level (Quercia, Ellis, Capra, & Crowcroft, 2012). Such techniques, though based on information about online communities, might allow the estimation of well-being in a geographic community as a complement to traditional survey methods, with potential impact on policy-making (Dodds, Harris, Kloumann, Bliss, & Danforth, 2011). In a large-scale study using Twitter data, Golder and Macy observed seasonal and diurnal patterns in mood that were correlated with work, sleep, and day length (Golder & Macy, 2011). Paul and Dredze found that Twitter activity correlated to public health metrics, demonstrating the potential for using social media in public health research (Paul & Dredze, 2011). Zagheni et al. have estimated international migration rates based on analysis of a Yahoo! email corpus (Zagheni & Weber, 2012) and data collected from Twitter (Zagheni et al., 2014).

Others have focused on individuals, considering how people's personal lives relate to online behavior. De Choudhury et al. have used mood patterns in the social media activities of individuals to understand behavior changes related to childbirth (De Choudhury, Counts, & Horvitz, 2013), and to recognize signs of depression (De Choudhury, Gamon, Counts, & Horvitz, 2013). This kind of approach could inform the creation of technology to identify depression and the development of better ways to support people suffering from depression.

In between these studies at the population and individual levels, some research considers offline and online phenomena in organizations or small communities. Communication tools used in organizations, such as email and chat, can generate insight about organizational structures and politics. Taking advantage of a rare opportunity to study an extensive organizational email record, Diesner et al., among others, have analyzed the dynamics of the communication network captured by the Enron email corpus in order to better understand the underlying causes of organizational

failures (Diesner et al., 2005). In this case, insights gained from the organizational email record may inform effective prevention and response to future crises.

As the above examples illustrate, message data from online social media and communication systems can provide insight about the relationship between online and offline activity at multiple levels, from the individual level all the way up to the societal level. The choice of a unit of analysis affects how social data researchers sample, select, and filter their data, and what methods and technologies are most appropriate. De Choudhury's studies on depression and life changes in social media took a mixed methods approach, using correlational analysis in combination with qualitative validation to illustrate and confirm the meaning of statistical findings. However, most of these large-scale studies used only quantitative methods and computational modeling, and did not incorporate qualitative approaches to examining the data.

### 2.1.2 Online Communities and Spaces

In contrast with the above work, which sees online social data as a lens for studying people's lives offline, other researchers have approached online spaces and online communities as social and cultural sites worthy of study in their own right.

Millions of people use online media and communication platforms to socialize and interact daily with friends, family, co-workers, and strangers. Records of online messages can provide insight about how people use online communication tools to support familiar social activities and goals, and how technology intersects with social practices. danah boyd's ethnographic studies of MySpace and other social network sites reveal how teens use social network sites as semi-private places for identity formation, status negotiation, and peer-to-peer sociality (boyd, 2007; boyd & Ellison, 2007). boyd et al. have also analyzed the practice of "retweeting" on Twitter as a

conversational practice, exploring issues of authorship, attribution, and fidelity (boyd et al., 2010). These works exemplify new "digital ethnography" methods (Murthy, 2008), combining extended online participant observation with interviews of online community members, through detailed interpretive analysis of people and social media content.

While still approaching online activity as their primary object of study, other researchers have focused on organization, structure, and emergent complexity in online communication and social media. In this space, mixed methods are commonly used. Goggins et al. use grounded theory analysis in combination with quantitative social network analysis to study the online collaboration of small groups in an online course, based on data from course forums, interviews with students, and field notes (Goggins et al., 2011); this methodological approach was later formulated as Group Informatics (Goggins, Mascaro, & Valetto, 2013). Also taking a grounded approach, Aragon et al. conduced qualitative analysis of thousands of chat messages and interviews to develop a theoretical understanding of creativity, emotion, and online collaboration spaces (Aragon, Poon, Monroy-Hernández, & Aragon, 2009).

Several researchers have studied online organization, structure, and information practices in the context of crisis events and protests. Agarwal et al. used a combination of manual coding of links shared on Twitter, ethnographic interviews, and large-scale social network analysis to study the formation and organizational behavior of the Occupy movement in New York City, Seattle, and Oakland (Agarwal, Bennett, Johnson, & Walker, 2014). Building on this approach, Bennet, Segerberg, and Walker developed theory about how production, curation, and integration processes support peer organization of large networked protests (Bennett, Segerberg, & Walker, 2014). Focusing on the 2011 Egyptian political uprisings, Starbird and Palen used a mixed methods analyses of thousands of tweets to explore the notion that Twitter communities accomplish

information processing work by collectively filtering and recommending content during crisis events (Starbird & Palen, 2012). To understand the spread of rumor and misinformation online during crises, Dailey and Starbird combined content analysis of social media activity with interviews and participant observation (Dailey & Starbird, 2014). In the context of the 2010 BP Deepwater Horizon oil spill, Starbird et al. conducted mixed methods analysis of Twitter posts, networks, and links to the outside internet, to understand how social media was used in a long-term environmental disaster (Starbird et al., 2015).

From these examples, it is evident that researchers in the social sciences have collected and analyzed social media and online communication data to address a wide variety of research questions. The specific approaches vary, but many researchers have found it useful to combine multiple data sources and multiple methods of analysis to answer their research questions. Whether comparing online activity to offline, or studying online social behavior itself, these studies are advancing fundamental social science questions, suggesting new ways of improving online communication platforms, and shaping policy and public expectations regarding online spaces.

## 2.2 Overview of Qualitative Research

The above studies often involve a high volume of social media and online communication data, which begs for the use of quantitative methods such as social network analysis, statistical analyses, and machine learning. However, many of the examples above also incorporated *qualitative* methods into their research design. I will discuss the specific data analysis practices of some of these researchers in greater detail in Chapter 3. In this section, I provide a brief history of the qualitative methodology and a discussion on practical qualitative data analysis techniques. The

purpose of this section is to provide the reader with a sense of the historical depth of qualitative traditions, and to outline some challenges in conducting qualitative research.

### 2.2.1 Historical Context

Qualitative research is a complex topic with many meanings in different circles, and a thorough review of the topic is well beyond the scope of this chapter. However, for a reader unfamiliar with qualitative research, the historical summary below may provide useful background and suggestions for further reading. This summary, abstracted from the more detailed historical account provided in the introduction to the SAGE *Handbook of Qualitative Research* (Denzin & Lincoln, 2011), focuses primarily on ethnographic approaches, which are at the core of qualitative research in sociology and anthropology.

Ethnography originated as early as the 15$^{th}$ and 16$^{th}$ centuries, in the descriptions and reports of indigenous peoples furnished by European explorers, missionaries, and colony administrators. The origin of anthropology and ethnography in a history of Western conquest and imperialism has cast a long shadow over the thinking and writing of qualitative researchers (Vidich & Lyman, 1993). Around the beginning of the 20$^{th}$ century, the practice of qualitative research became a professional discipline located in the fields of anthropology and sociology. Early qualitative research was concerned with the objective capture and description of the cultures of the strange and alien "Other," exemplified by anthropologists such as Bronisław Malinowsky and Margaret Mead.

After World War II, building on the legacy of traditional ethnography, *modernist* qualitative research focused on formalizing methods to produce more "rigorous" studies. Qualitative researchers borrowed concepts from quantitative research, such as internal and external validity, and began calculating "quasi-statistics" based on ethnographic data. Modernist qualitative research

is based on a postpositivist epistemology that sees our ability to observe the world as inherently limited. It often employs multiple methods to obtain a complete and reliable account. Representative of this period is Glaser and Strauss's *The Discovery of Grounded Theory* (Glaser & Strauss, 1967), which continues to see widespread use today outside its originating field of sociology (Charmaz, 2006; Strauss & Corbin, 1990). The *objectivism* reflected in these early works has been linked to the field's colonial heritage, and some of the assumptions of these early ethnographies are considered problematic in contemporary qualitative research: "Ethnographies do not produce timeless truths. The commitment to objectivism is now in doubt. The complicity with imperialism is openly challenged today" (Denzin & Lincoln, 2011).

After 1970, Denzin and Lincoln describe a "blurring" of qualitative research genres. A wide range of methods and approaches were employed, including, for example, symbolic interactionism, constructivism, phenomenology, ethnomethodology, critical theory, semiotics, and feminism. (Denzin & Lincoln, 2011, p. 23). The available ways of collecting, analyzing, and reporting data expanded considerably. Anthropologist Clifford Geertz focused attention on *interpretation* and "thick description" of particular events and customs (Geertz, 1973). Researchers reconsidered the politics and ethics of their work, and questioned the author's role and position in their texts. In recent decades, the discipline of qualitative research has faced crises: it is no longer accepted that the ethnographic researcher directly captures the "lived experience" of its subjects; rather, experience is thought to be created, in a sense, through the act of writing (Denzin & Lincoln, 2011).

Qualitative research methods are practiced in different ways in many different communities, and most of these different historical styles of qualitative research remain in active use today. Many of the social scientists whose research is discussed in this dissertation borrow elements from one or more of the styles described above, but incorporate these elements into their own research in a

pragmatic fashion. This capacity for adaptation and reformulation based on the needs of particular research sites, sometimes referred to as *bricolage*, is considered characteristic of qualitative research (Denzin & Lincoln, 2011). This flexibility also makes qualitative research difficult to define. Denzin and Lincoln provide the following definition:

> Qualitative research is a situated activity that locates the observer in the world. It consists of a set of interpretive, material practices that make the world visible. These practices transform the world. They turn the world into a series of representations, including field notes, interviews, conversations, photographs, recordings, and memos to the self. At this level, qualitative research involves an interpretive, naturalistic approach to the world. This means that qualitative researchers study things in their natural settings, attempting to make sense of, or interpret, phenomena in terms of the meanings people bring to them.

Although a variety of approaches are labeled qualitative research in various disciplines, this definition raises a number of key issues which are essential to contemporary qualitative research. Bogdan and Biklen summarize qualitative research as naturalistic, using descriptive data, concerned with process, inductive, and meaning focused (Bogdan & Biklen, 1997). Qualitative research emphasizes situated activity, outside a lab. It depends on interpretation and meaning-making by the researcher, the reader, and the people being observed. It addresses the observers' relationships and interactions with the world, employs rich, descriptive media as data, and is presented in a format which communicates the richness of the data (e.g. narratives, quotations).

### 2.2.2 Practicalities of Qualitative Research

Despite the philosophical aspects of the historical account above, the day-to-day practice of qualitative research tends to be down-to-earth and pragmatic. Researchers are encouraged to adopt, adapt, and invent methods of data collection, methods of analysis, theoretical stances, and reporting techniques to best serve their research objectives (Denzin & Lincoln, 2011). Thus, qualitative researchers often use different methods to gather multiple kinds of data and analyze it in multiple ways. Nevertheless, this section discusses the traditional practices of ethnographic data collection and analysis, which are common across many variants of qualitative research.

Because of the richness of data that qualitative researchers seek to obtain, data collection is usually a time consuming process requiring extensive time in the field and substantial rapport with informants. Two classic methods of qualitative data collection are *participant observation* and *interviewing* (Bogdan & Biklen, 1997). In participant observation, the ethnographer goes out into the world, becomes part of the setting that they are studying, forms relationships, and observes what is going on in context, hence, becoming a *participant* observer (Emerson, Fretz, & Shaw, 2011). The researcher systematically makes detailed ethnographic *field notes* to capture their observations, creating an "accumulating written record of these observations and experiences" (Emerson et al., 2011). Some researchers also record *memos*, separate documents which capture their own impressions, musings, and ideas. Memos are a useful way to condense a heap of field notes into findings. Researchers also use memos to reflect on their own place and role in the research site. Participant observation is usually a relatively open-ended research activity; the observer remains open to the unexpected and prefers to record events of interest in detail, rather than to prematurely restrict the research focus.

Many qualitative studies combine participant observation with interviewing. In contrast to structured interview or survey procedures, ethnographic interviews are often *semi-structured* or *unstructured*, meaning that the questions asked are flexible and may be altered, reordered, extended, or abandoned on the spot by the interviewer based on the contingencies of the situation (Weiss, 1995). While this means that the answers given by interviewees are not comparable to one another in the way that measurements obtained from experimental subjects are comparable, the researcher is able to obtain more detailed explanations, examples, and stories which are necessary to understand the perspectives, thinking, motivations, and meanings of the interviewees (Weiss, 1995). Qualitative interviewing tends to use small focused samples, and the selection of informants to interview is not typically based on the ideal of random sampling. Instead, it is purposeful, based on situational or theoretically-driven concerns. Interviews are usually recorded and then transcribed, and it is these transcripts which constitute interview data.

The *analysis* of field notes and interview transcripts, as well as any other data that have been collected (e.g. photos, video), is a contentious topic within qualitative research. Analytical practices differ in their formalism, epistemological rationale, and historical origin, and are employed with varying degrees of faithfulness; often qualitative researchers depart from established practice to better serve the needs of the current research. Recognizing this diversity, I will focus on *grounded theory* as an example (Glaser & Strauss, 1967); the grounded theory family of methods has achieved widespread adoption in the social sciences and, in particular, has been adapted for mixed methods studies with online social data (Goggins et al., 2013; T. J. Scott et al., 2012; Starbird et al., 2015). Grounded theory is used to inductively build theory based on rich qualitative data, with strong emphasis on the emergence of theoretical constructs from the data, rather than from preexisting beliefs or outside sources.

From a practical standpoint, grounded theory relies on two key activities: *coding* and *memoing*. Coding refers to an *analytical* process where the researcher reads through the collected data attaching *codes* (i.e. categorical labels) to sections of the text: "Coding distills data, sorts them, and gives us a handle for making comparisons with other segments of data" (Charmaz, 2006). In a grounded theory process, the researcher begins with *open* coding, where codes are derived from the data, and transitions iteratively towards focused, theoretically-structured coding. Memos are written during data analysis as well; the writing of memos is a *synthetic* process which gradually builds and develops larger, higher-level categories, relationships, and theories based on ("grounded in") the data.

Grounded theory's codification of analytical procedures provides a clear route to making sense of qualitative data; this is one reason that grounded theory methodologies are so widely used (Charmaz, 2006). The practical formalism of grounded theory helps qualitative researchers explain and defend their approach in communities where quantitative and experimental research is the norm. Furthermore, researchers in many fields have combined grounded theory-style qualitative analysis with quantitative techniques (e.g. descriptive statistics) in mixed methods research, while maintaining, as their overall approach, grounded theory's emphasis on inductively constructing knowledge based on empirical data (Bergman, 2008; Charmaz & Belgrave, 2002; Creswell, 2014). Thus, while grounded theory traditionally was developed as a specific procedure for qualitative data analysis, today it is often used to describe a wide range of practices incorporating aspects of the original grounded theory method.

Iteratively coding and memoing over reams of interview transcripts and observational data is a time-consuming, labor-intensive process. There is ample opportunity for error, losing track of data, or being unable to locate the right data at the right time. With the wide availability of computers,

qualitative research has also transitioned to digital data, and a variety of software packages have been created, such as ATLAS.ti and NVivo. Coding and analyzing data using specialized software can enable many efficiency and validity improvements over traditional paper and pencil techniques, and these packages are in wide use. However, exemplifying the self-reflection characteristic of the qualitative research perspective, some researchers have cautioned against an unquestioning adoption of software, which may subtly constrain or inadvertently damage the research process (Goble, Austin, Larsen, Kreitzer, & Brintnell, 2012; St John & Johnson, 2000). While software tools for quantitative data analysis have been accepted and used widely for decades, the interpretive flexibility of qualitative methodology demands that researchers carefully consider how the introduction of qualitative data analysis software impacts their research questions, processes, and findings. Tools for qualitative analysis are discussed in Chapter 6.

## 2.3 Overview of Machine Learning

The daunting prospect of qualitatively analyzing large online social datasets, which often contain thousands or millions of short text-based messages, suggests opportunities for automation. Machine learning techniques, such as text clustering and text classification, can be used to reveal and create structure in social media and online communication datasets, and have occasionally been applied in qualitative or mixed methods approaches (Crowston, Liu, & Allen, 2010; Janasik et al., 2009; Rosé et al., 2008). The applicability of machine learning for automatically coding online communication text is discussed in greater detail in Chapter 4.

In this section, I summarize some background information on machine learning terminology and concepts, as well as standard methods for evaluation of classification algorithms. While a thorough overview is beyond the scope of this chapter, my goal here to provide readers unfamiliar with the

subject a sense of the vocabulary and philosophical perspectives of machine learning and other computational methodologies, and to suggest directions for further reading.

### 2.3.1 Machine Learning Concepts and Terms

The purpose of this section is to introduce concepts and vocabulary associated with machine learning and data mining. The summary below draws from the textbook *Data Mining: Practical Machine Learning Tools and Techniques* by Ian H. Witten and Eibe Frank (Witten, Frank, & Hall, 2011), two creators of the popular open-source *Weka* machine learning toolkit.

Machine learning is the use of various automated techniques to build useful models of data. As described by Witten and Frank, *machine learning* is "the acquisition of structural descriptions from examples." Machine learning starts with a list of *examples* (sometimes called instances), usually represented by a table of rows and columns; rows represent independent examples, and the columns contain different kinds of abstract information about those examples, called features or attributes. This collection of examples is called *training data*. A computer reads the training data and attempts to detect and describe a useful pattern. The specific technique used to accomplish this is called a *machine learning algorithm*, and a great many have been invented (e.g. Naïve Bayes, logistic regression, ID3). Each algorithm is capable of learning only certain kinds of patterns from the data. The pattern learned is called a *model*, and may be represented in different ways (e.g. linear separator, decision tree, list of rules). Subsequently, when new examples are encountered, the learned model can be used to evaluate the example and make some kind of prediction about it.

The models which are learned from the training data are deemed useful when they can be productively applied to new unforeseen examples in the future. Improving the performance of the

model may require changing or adjusting the algorithm, developing more informative *features* (attributes of the data which are available to the machine) (S. Scott & Matwin, 1999), gathering more training data (Domingos, 2012), or cleaning and improving the data (Rosé et al., 2008). The field of machine learning in computer science is concerned with developing and evaluating new machine learning algorithms, understanding the general problem of learning in new and useful ways, and exploring applications of machine learning. It has strong connections to both artificial intelligence and statistics.

There are several important categories of machine learning. First, machine learning can be organized according to the type of problem the model is intended to solve. One type of problem is *classification*. In classification problems, the model is used to classify an example according to a fixed set of categories, or *classes*. For example, a researcher might use a machine learning model to classify tweets into several emotion categories (Purver & Battersby, 2012). For classifying text-based data, the sub-field of *text classification* has developed a large number of useful techniques (Sebastiani, 2002). Another type of machine learning problem is *regression*, which involves using a learned model to predict a continuous value based on examples, instead of predicting from a fixed set of classes.

Machine learning techniques can also be organized based on how they are set up. In *supervised* machine learning, the machine learning algorithm is given training data that includes the desired or expected *ground truth* output, perhaps created manually beforehand, by an expert. The algorithm extracts a model which hopefully reflects the ground truth labels. For example, consider the "iris" dataset, a classic dataset describing flowers; there, the input data includes measurements from 50 different types of iris flowers: sepal length, sepal width, petal length, and petal width (Witten et al., 2011). The training data also includes the flower species of each example, as ground

truth. A classifier trained on this data could be supplied with an unlabeled example flower, from which we hope it can guess the flower's correct species.

In many practical examples, creating ground truth data is expensive and time consuming. In *unsupervised* machine learning, models are built from data that does *not* include ground truth labels. One common unsupervised technique is *clustering*, where data are automatically organized into groups or clusters according to some notion of similarity or distance between examples. Janasik et al. explored the use of clustering with self-organizing maps to improve results in a qualitative research project (Janasik et al., 2009), while Prier et al. have used unsupervised topic-modeling to explore tweets about smoking (Prier, Smith, Giraud-Carrier, & Hanson, 2011). Unsupervised learning is often used as a technique for exploratory data analysis, where it is considered successful when it reveals some structure in the data. However, because ground truth data is not available in this setting, unsupervised learning is difficult to evaluate. The discussion below regarding evaluation focuses on supervised classification.

### 2.3.2 Evaluating Classification Algorithms

Evaluation is a crucial step in machine learning, because there numerous biases that can interfere with the usefulness of learned models. This section outlines evaluation problems in context of using machine learning to automatically categorize data (i.e. classification), where priorities and standards for evaluation are relatively clear.

When training a machine learning model, the developer needs to be able to assess the quality of the pattern which has been learned from the training data. This is typically done in terms of statistical performance: how accurate do we think the model's output will be when faced with data in the future? Predicting future performance is not straightforward. While a given model might

appear to accurately model the *training* data, performance on new data is often much worse. The training data may contain sampling biases as well as random noise and errors. Both systematic bias and noise in the training data can interfere with machine learning. This can produce a common problem described as *overfitting*, or failure to generalize: the model may learn patterns from the training data which do not exist in future data (Domingos, 2012).

A typical solution to this dilemma is to employ a two-step process of training followed by testing. The labeled data is divided, randomly, into a subset for training and a subset for testing. Once the model has been created from the training portion of the data, the model is used to evaluate the test data; its predictions on the test data are compared to the ground truth labels to estimate how well the model might generalize to data on which it was not trained. These estimates can still be threatened by sampling bias and other issues, and in practice more sophisticated procedures are typically used. Often the train/test procedure is repeated several times with different train/test splits, so that the influence of random sample noise can be minimized.

When a model's predictions are compared against ground truth data, assessment of quality relies on various statistical metrics. One simple metric is success rate or accuracy (i.e. percent of the data classified correctly), or equivalently, the error rate. No metric is perfect: there are several different kind of mistakes that trained classifiers can commit (e.g. false positives and false negatives), and the success rate lumps these errors together. For comparing classifiers and for debugging, it is sometimes important to unpack these different types of errors in detail. Other metrics often used include rates of true positives and negatives, precision and recall (Raghavan, Madani, & Jones, 2005), and receiver operating characteristics (Bradley, 1997). The selection of an appropriate metric must account for the intended application, as optimizing some metrics will incur tradeoffs in others. For example, in a spam filtering application, false positives should be minimized so that

non-spam emails do not get rejected. In a cancer screening application, where the cost of failing to detect the disease is very high, it may be more valuable to minimize false negatives.

### 2.3.3 Technical Challenges

Machine learning can be a powerful tool for exploring large datasets, and automatic clustering and classification can be applied to support qualitative investigations. However, using machine learning successfully in practice is challenging, as considerable expertise is required to avoid various pitfalls and to achieve good results. There is a need for better machine learning tools, not only for qualitative researchers and social scientists, but also for machine learning developers.

Based on the idea that allowing users to interact with machine learning algorithms can be useful (Stumpf et al., 2007), interactive machine learning researchers have developed visualization techniques for understanding and debugging model performance more quickly (Amershi et al., 2015; Talbot, Lee, Kapoor, & Tan, 2009; Torkildson, 2013), interaction paradigms and principles for steering classifiers (Kulesza et al., 2009; Kulesza, Amershi, Caruana, Fisher, & Charles, 2014) and systems for discovering and developing better features (Brooks et al., 2015; Heimerl, Jochim, Koch, & Ertl, 2012; Raghavan et al., 2005). Others have explored the machine learning development process holistically, creating integrated development environments to manage training, exploration, debugging, and evaluation (Patel, Fogarty, Landay, & Harrison, 2008; Simard et al., 2014). Much work remains to be done in this space, and few of these advances have yet entered the market as usable tools, but the potential benefits of human-centered machine learning approaches are intriguing. Implications for machine learning in qualitative and mixed methods social science research with online social data are discussed in Chapter 4.

# Chapter 3: Social Media Data Analysis Practices

Researchers working with social media and online communication data face many hurdles, both methodological and technical. How should research be done with online social data, to ensure that the work meets quality standards, e.g. for validity, ethics, and reproducibility? What skills and tools are needed to make high quality research possible? As social media data becomes more widely used in the social sciences, these questions are part of ongoing discussion within the research community (boyd & Crawford, 2012; Crowston & Nahon, 2015).

There are a variety of general-purpose tools used for social media data analysis, such as Microsoft Excel, Google Spreadsheets, Tableau, and Gephi; researchers also use custom scripts and programs. However, incompatibilities and design limitations in these tools require researchers to perform expensive transformations, brittle workarounds, and context shifts that hinder their progress. Yet, if technology designers and researchers are to understand how to develop better analytical tools for social scientists, there are many unanswered questions. What kinds of tools and functionalities are needed? What types and structures of information should they work with? What analytical tasks and activities should they support? In human-centered design, good design depends on having a deep understanding of users (Norman, 2013). While researchers in visual analytics and HCI have conducted studies to understand data analysis practices in several domains, including intelligence analysis (Chin, Kuchar, & Wolf, 2009; Kang & Stasko, 2011), building design (Tory & Staub-French, 2008), automotive engineering (Sedlmair et al., 2011), and enterprise data analytics (Kandel, Paepcke, Hellerstein, & Heer, 2012), there has been little research on the data analysis practices of social scientists working with social media data.

In this chapter, my goal is to develop a grounded understanding of social data analysis practices, and to draw out implications for the design of tools and software that could support mixed methods analysis of social data. I ask the following question: *What are the goals, barriers, tools, and processes associated with mixed methods online social data research?*

To answer this question, my colleagues and I have interviewed several individuals who are conducting research with social media and online communication data. My informants in this study are all working intimately with social media data to address social science questions. Many are drawing on both qualitative and quantitative analysis techniques, and actively exploring new methods and approaches to analyzing their data. The rich description of work practices obtained by interviewing these researchers suggests implications for the design of new data analysis tools for social media data.

This chapter describes my research method and participants, and then explores the social science research practices that I observed during my study. The findings are organized around data collection practices, the structure and character of social media datasets, and analytical processes, methods, and techniques. Throughout, I focus on challenges that researchers face in gathering and working with their data. I conclude by discussing implications for the design of social media data analysis tools.

## 3.1 Acknowledgments

Thank you to the researchers who agreed to be interviewed and observed for this chapter. Three of the interviews were conducted in cooperation with colleagues Ray Hong, Sanny Lin, Zening Qu, Jeff Smith, and Rafal Kocielnik as part of a collaborative research group. The remaining interviews and observations, data analysis, and writing were completed by the author.

## 3.2 Related Work

In this section, I examine background and related work pertaining to social media data analysis practices, research on the practice of scientists and data analysts, and research on new social science methodologies for social media.

### 3.2.1 Data Analysis Practices

The challenge of designing data analysis tools is inherently interdisciplinary; it requires some understanding of not only the domain in which the data analyst works, but also of numerous technical fields. While not all data analysis tools are visual, the interdisciplinary field of visual analytics has made significant progress towards understanding how to support data analysis through technology that integrates interactive visualizations and computational analytics. This section discusses the role of work practice studies in designing data analysis tools.

In their reflection on the history and role of workplace studies in the field of computer-supported cooperative work (CSCW), Luff, Hindmarsh, and Heath argue that many early efforts to develop technology supporting collaboration failed because of poor understanding of how actual collaboration takes place in the real world (Luff et al., 2000). Following from some of those early failures, the CSCW community began conducting more naturalistic studies aimed at understanding how works takes place in practice. Although such naturalistic studies of work practice have long been an essential component of the human-centered design toolkit (Norman, 2013), a similar trend can be seen in visualization and visual analytics. Visualization and visual analytics researchers have pointed to a lack of uptake and adoption of their techniques and tools in the real world, which they see as evidence that greater attention should be given to understanding their users and the context in which data analysis work takes place (Munzner, 2009; Plaisant, 2004). In the visual

analytics community, this has taken the form of calls for naturalistic field-based work practice studies of specific application domains, drawing on ethnographic observation and interview methods, with an eye towards design implications (Munzner, 2009).

One domain where this call for design-oriented research on work practice has been answered is that of intelligence analysis, historically one of the primary application areas for visual analytics technology (Thomas & Cook, 2005). Intelligence analysis involves making sense of a great quantity of complex, changing, ambiguous data in a time-sensitive, politically charged context, and thus, it is a particularly challenging case for technology design. Several qualitative studies of intelligence analysis work practices have been conducted with the goal of informing the design of better technology to support intelligence analysts, uncovering and describing various surprising, counterintuitive, but crucial aspects of work practice. Based on qualitative observation of analysts workshopping a set of prepared scenarios, Chin, Kuchar, and Wolf described analytical strategies and processes, tools and software, perspectives on credibility, and collaboration practices; the authors propose design implications such as support for systematizing standard analysis perspectives, combining multiple large screens, incorporating affordances from traditional physical media (e.g. highlighters and pens), supporting link analysis, and integrating historical information (Chin et al., 2009).

Taking a longitudinal approach, Kang and Stasko observed teams of intelligence analysis students over a ten week period (Kang & Stasko, 2011). Comparing their observations with prevailing assumptions and preconceptions, Kang and Stasko found that actual practice was organic and parallel, not sequential as in sensemaking (Pirolli & Card, 2005); significant time was spent in the "framing" phase of sensemaking (Klein, Moon, & Hoffman, 2006), rather than analyzing specific sets of data; collaboration was common-place; and the dominant need is not sophisticated

technology to support narrow analytical techniques and processes, but rather support for managing the overall process more effectively, and for efficiently working with an ecosystem of many tools and techniques (Kang & Stasko, 2011).

Similar formative naturalistic studies have been performed in a few other visual analytics application domains, e.g. building design (Tory & Staub-French, 2008), automotive engineering (Sedlmair et al., 2011), and enterprise data analytics (Kandel et al., 2012), but many researchers in visualization and visual analytics continue to focus on innovative designs and techniques without solid grounding in current practices. One reason for this is that obtaining deep characterization of a domain is a challenging and costly endeavor (Brehmer, Carpendale, Lee, & Tory, 2014). However, it is important to note that, as has been found in CSCW and HCI, work practice studies in specific domains have implications beyond the immediate context of the study, both revealing broader patterns and contributing to theory that is often applicable across domains (Luff et al., 2000; Orlikowski, 2000; Schmidt, 2000).

While little attention has been paid to understanding practices, researchers have designed and evaluated a number of tools for analyzing and exploring social media data, such as the "Visual Backchannel" tool for monitoring events on Twitter (Dork, Gruen, Williamson, & Carpendale, 2010); *Vox Civitas*, a tool for journalists using Twitter as a source (Diakopoulos, Naaman, & Kivran-Swaine, 2010); and twitInfo (Marcus et al., 2011). Some of these tools are designed with a deep understanding of their target domains (e.g. journalism, event management, business marketing), but even here, few have directly addressed the needs of *social scientists* working with social media data. Sawyer et al. have discussed distributed work practices in the social sciences, but did not address online social data analysis or design implications for data analysis software

(Sawyer, Kaziunas, & Østerlund, 2012). With this chapter, I aim to fill that gap, inform technology design, and suggest questions for future research.

### 3.2.2 New Methodologies for Social Media Data

Faced with new types of data and new challenges, researchers are developing new methods and publishing articles on the methods they are exploring. These researchers are asking how traditional research methods can be adapted or "scaled up" to online social data, and what new methods might be developed to allow researchers to study social media data in new ways.

For example, informed by social network analysis, information science, and theories of small group cooperation, Goggins et al. have developed a methodological approach known as "Group Informatics" (Goggins et al., 2013), to help researchers study the social connections between members of technologically mediated groups. The Group Informatics framework affords a type of mixed qualitative and quantitative social network analysis, and involves a grounded-theory-inspired process of classifying and weighting small-group interactions, followed by the construction and analysis of a weighted social interaction network that can be used for comparable, grounded, network analysis (Goggins et al., 2011). In this way, it manages to fuse qualitative and quantitative approaches into a unified methodological framework.

While not always adopting such a formal approach to methods development, other researchers also draw on both qualitative and quantitative methods to analyze social media data. De Choudhury et al. used machine learning and qualitative analysis to analyze signs of depression in Twitter (De Choudhury, Counts, et al., 2013). Starbird et al. deployed qualitative analysis of social media content, also inspired by grounded theory, in combination with statistical and visual analysis to characterize information practices on social media during major crises (Maddock et al., 2015;

Starbird et al., 2015; Starbird & Palen, 2012). Rosé et al. used structured qualitative coding to analyze learning processes in a computer-supported collaborative learning forum, and then apply machine learning technology to automatically code new corpora (Rosé et al., 2008). The motivation for doing so is primarily to reduce cost and increase efficiency, but Rosé et al. discuss the benefits and limitations of this approach in greater depth.

The flexibility, adaptability, and open-ended character of these approaches affords descriptive power, given the challenges presented by the complex structures and meanings in social media datasets. In the body of published social media and online communication research, the explosion of methodological innovation is a positive step towards finding more powerful and accessible ways of working with social media and online communication data in the social sciences.

## 3.3 Method

The goal of this study is to inform the design of data analytics systems for social scientists working with social media data. By focusing attention on the everyday practices of social scientists, it is possible to learn how they structure their interactions with technology to accomplish their goals (Orlikowski, 2000). Ethnographic research methods are effective for discovering the constraints, artifacts, and meanings in everyday work practice, and are often used to develop implications and requirements for design (Luff et al., 2000). In this chapter, I primarily use data from semi-structured interviews (Seidman, 2005; Weiss, 1995), along with a small amount of formal field observation (Emerson et al., 2011).

I conducted these interviews after two years of actively collaborating in social media and online communication research projects (discussed in Chapters 4, 5, and 6). Those experiences provided experiential background knowledge that I drew on in this study when selecting research questions,

recruiting participants, and interpreting my ethnographic data. While it is more typical, in human-centered design methodology, for this kind of study to precede and inform later design and development work, here the ethnographic research was conducted afterwards. Therefore, it complements and contextualizes, but does not directly inform, the design projects discussed in the following chapters. I include it first in this dissertation because, for the reader, it effectively motivates and grounds the subsequent chapters in social scientists' work practices.

### 3.3.1 Research Site and Recruitment

While researchers studying social media and online communication are scattered around the world, the University of Washington hosts significant activity in this area. Faculty hiring over the past few years, intense student interest, and broader recognition and support for data-intensive research across campus have led to several new research labs and groups, new learning and knowledge-sharing events, and the gradual knitting together of a university-wide community of social media and online communication researchers.

In this environment, I recruited researchers who were working with social media and online communications data, especially larger amounts of data: hundreds of thousands or millions of messages. These larger datasets introduce more acute technical and methodological challenges than small and medium-sized datasets, and so I expected to find more pronounced opportunities and needs for technology in these contexts. I focused on researchers working with *content* in their data. Researchers might typically consider social media and online communication from a variety of perspectives, e.g. they may look at interactions, network structures, change and dynamics, opinion and sentiment, information and flow. I sought researchers who were asking questions about what social media users communicate *about*, i.e. researchers who take into account the

meaning of the content they study. This aspect of the data presents formidable technical and methodological challenges to designers and researchers: how can researchers effectively produce knowledge from the text of a million disparate messages? Few researchers consider only one aspect of their data, but my informants had a real interest in the text-based content of their datasets.

Because UW is home to several researchers working with social media and online communication data, I easily made contact with potential informants. The research group discussed in Chapter 4 allowed me to become an active participant in the local community studying online communication and social media. Connections with colleagues from that project and other research collaborators put me in touch with many of my informants. I attended community learning events and workshops, and through those events, I met and recruited researchers across multiple labs and departments studying social media.

Several of my informants have been collaborating with one another in research groups and projects, and as such, their practices are aligned in many ways; they share some research techniques and approaches. Because they are part of the same community at the University of Washington, they have commonalities that may not be shared by researchers studying social media and online communication data at other sites. It is important to keep this bias in mind, but this limitation comes with a corresponding strength: the deep, concrete, and descriptive understanding of social media research practices that comes from a narrowly focused ethnographic study may be more informative for design decisions (Blomberg, Giacomi, Mosher, & Swenton-Wall, 1993). Moreover, due to the intensely interdisciplinary and collaborative work that my informants typically engage in, considering them in isolation of their communities and collaborators would distort what we can learn about their practices.

### 3.3.2 Scope of Interviews and Observations

I conducted interviews with seven researchers studying social media and online communication at the University of Washington between October 2014 and June 2015. Interviews were semi-structured, and the questions were targeted to the specific situation of each interviewee. Interviewees were asked to focus on a recent research project and to discuss the following:

- Research goals and objectives, phases and timeline, questions
- Backgrounds and roles of interviewee and collaborators
- Data, including origins, processing, format, and flow
- Data analysis processes, methodology, research design decisions
- Problems and challenges faced, breakthroughs and successes
- Data analysis, processing, and visualization tools used

A few of the seven interviewees were interviewed more than once, and overall eleven hours of recordings were obtained. These were transcribed using an online transcription service[1], and then checked manually for transcription errors.

In addition to interviews, I observed several hours of research group meetings. In these meetings, I took notes by hand, which I later expanded into detailed descriptive ethnographic field notes (Emerson et al., 2011). My note-taking recorded much of what was said in the group meetings, but I particularly paid attention to discussions about challenges and difficulties, as well as research goals and questions. Because the diverse interests and broad scope of the research group I observed, only part of these observations were related to social media and online communication

[1] Rev.com

data. Therefore, I have included notes from only one hour of these meetings in this chapter. During this particular meeting, one group member, Anna, gave a presentation about her work on social media data collection and analysis for sociological studies.

Table 3.1 provides summary information on the participants. Pseudonyms are used to protect anonymity. One participant, Shawn Walker, requested that his real name be used.

| Name | Field | Level | Project | Data Sources |
|---|---|---|---|---|
| Jamie | Environmental Studies | PhD student | Assessing public opinion on Twitter. | Twitter |
| Alan | Communication | PhD student | Comparing social media platforms. | Twitter, Sina Weibo |
| Karen | Information Science | PhD student | Information practices online and offline. | Twitter, interviews |
| Laura | Sociology | Faculty | Using social media to study informal communication. | Twitter, other disaster data |
| Eric | Human-Computer Interaction | Undergrad. student | Spread of information in social media. | Twitter, news media |
| Leo | Sociology | Faculty | Studying global patterns using social media. | Twitter, Google+, Facebook |
| Shawn | Information Science | PhD student | How to collect/manage social media data for research. | Twitter |
| Anna | Sociology | PhD student | Demographics in social media. | Twitter |

**Table 3.1: Interview and observation participant details. Names given are pseudonyms.**

### 3.3.3 Data Analysis

I conducted exploratory thematic analysis of the interview transcripts and field notes using a grounded theory approach (Charmaz, 2006; Glaser & Strauss, 1967). I applied open coding to the interview transcripts and field notes, looking for references to process, workflow, technology, challenges, and data analysis. While performing open-ended coding, I wrote memos documenting and developing the emerging themes. After the initial phase of coding, which resulted in a set of

over 80 codes, I used affinity diagramming to group the codes into a higher-level schema. This resulted in the following 8 categories:

- Analysis: techniques for revealing, breaking down, and restructuring data (e.g. reading, clustering, coding).
- Aspects: facets of social media data that were of interest to informants (e.g. time, content, networks).
- Challenges: problems and issues facing informants (e.g. data size and structure, irrelevance, privacy and ethics, access to data).
- Goals: purpose and desired outcomes of research activities (e.g. social science questions, tools, methodology).
- Methods: overall approaches and methodologies used by informants (e.g. grounded theory, content analysis, computational methods, mixed methods).
- Process: research processes/phases (e.g. exploration, iteration, transformation, integration).
- Skills: what informants know, how they learn it, and how their knowledge intermingles in collaborative projects.
- Tools: different kinds of tools (e.g. data collection & storage, visualization, collaboration) as well as infrastructural and tool ecology issues.

For these groups, I wrote memos reflecting on the group's meaning in light of the interview data. I focused on the Process, Analysis, and Challenges codes for deeper elaboration. Process helps to frame the work that participants are doing, while examples of analytical techniques point towards implications for design and were discussed in some depth. Many of the challenges discussed by informants are well-known, but they highlight areas for discussion, future work and, potentially,

design opportunities. I examined the data coded in these areas and began developing a theoretical organization for the findings, including supporting quotations and examples. With this focus in mind, I continued coding, iterating on the findings and implications, discussed below.

## 3.4 Findings and Discussion

In this section, I discuss the findings from my analysis. I begin with data collection, usually the first hurdle faced by social scientists who wish to study social media data. I then turn to technical and data processing challenges stemming from data structure and quality issues. Next, I discuss relevance and exploration, analytical tools and software, and several key data analysis techniques. Finally, I touch on socio-organizational challenges such as privacy and ethics.

### 3.4.1 Collecting Social Media Data

Researchers working with social media data dedicate a great deal of time, attention, and effort towards data collection. While collecting social media data may be less costly in some ways than traditional social science research methods (e.g. surveys, interviews, etc.), it can be more complex than it seems at first. For background, I will outline how data is obtained from the Twitter API[2], used by all of my participants. The Twitter API provides a set of commands which are called over the web; these commands respond with payloads of Twitter data, e.g. information about tweets or user accounts. An example request and response from the API documentation is in Figure 3.1.

Although some out-of-the-box solutions are available, most researchers in my study collect data by writing small scripts or programs that run on the researcher's computer; these programs make web requests to the Twitter API endpoints, receive payloads of data as a response, and save the

[2] https://dev.twitter.com

results in a file or database. Writing these programs typically involves learning the details of the Twitter API endpoints, options, and data format (which participants reported can include over 200 different fields). Some researchers use software libraries that abstract the details of the API and simplify this process, but the learning curve is still steep.

**Raw API Request**

```
GET https://api.twitter.com/1.1/statuses/
       user_timeline.json?screen_name=twitterapi&count=1
```

**Python Sample**

Using the tweepy[3] library

```
import tweepy
auth = tweepy.OAuthHandler("CONSUMER_KEY", "CONSUMER_SECRET")
auth.set_access_token("ACCESS_KEY", "ACCESS_SECRET")
api = tweepy.API(auth)
for tweet in tweepy.Cursor(api.user_timeline,
                           screen_name='twitterapi', count=1).items():
    print tweet
```

**Response**

tweet by the *@TwitterAPI* account

JSON format

Some fields omitted for brevity

```
{
  "created_at": "Wed Aug 29 17:12:58 +0000 2012",
  "contributors": null,
  "text": "Introducing the Twitter Certified Products Program:
https://t.co/MjJ8xAnT",
  "retweet_count": 121,
  "id": 240859602684612608,
  "retweeted": false,
  "in_reply_to_user_id": null,
  "user": {
    "name": "Twitter API",
    "created_at": "Wed May 23 06:01:13 +0000 2007",
    "location": "San Francisco, CA",
    "favourites_count": 24,
    "utc_offset": -28800,
    "followers_count": 1212864,
    "time_zone": "Pacific Time (US & Canada)",
    "description": "The Real Twitter API. I tweet about API changes,
service issues and happily answer questions about Twitter and our API.
Don't get an answer? It's on my website.",
    "statuses_count": 3333,
    "screen_name": "twitterapi"
    ...
  }
  ...
}
```

**Figure 3.1: Sample Twitter API request for tweets by a user account. The same request is shown in Python, along with the API response (one tweet in JSON format).**

[3] http://www.tweepy.org/

Researchers commonly save the data delivered by the Twitter API either in flat files or databases. Two file formats were commonly used by my interviewees: JavaScript Object Notation[4] (JSON), and Comma-Separated Values[5] (CSV). JSON is the native format in which data is provided by Twitter, illustrated in Figure 3.1, and can be easily taken in by popular programming languages. CSV is a tabular data format that is compatible with popular spreadsheet software, easier for humans to read, and more space efficient than JSON. Most of my interviewees convert portions of their datasets to CSV for some stages of analysis.

Aside from flat files, researchers also use databases to save social media data. The dynamic search, filtering, and grouping functions of database engines can be very useful for data analysis; these operations are typically much more difficult to run on flat files. On the other hand, researchers have to learn a new set of skills to set up, maintain, and use a database engine, which is a barrier for many. Interviewees reported using both traditional relational databases (e.g. MySQL), which store data in a normalized, indexed tabular format, and NoSQL databases (e.g. MongoDB), which store data in an object format similar to the native JSON provided by Twitter.

Because the data are owned by social media companies, researchers are at the mercy of corporations for access to quality data. Twitter stands out by having a public API that provides reliable access to a portion of its data, but few companies have this level of access. Even here, though, Twitter and other companies design their APIs to support commercial scenarios, rather than researchers. As a result, the details of these APIs sometimes create inconvenient constraints

[4] http://json.org/

[5] https://tools.ietf.org/html/rfc4180

that researchers must work around. Twitter and other social media companies also change how their APIs work, requiring researchers to update and maintain their software to keep up.

#### 3.4.1.1 Barriers to Access

For social scientists, there are many barriers to accessing social media data. Researchers typically write their own scripts to connect to APIs and download data, and learning programming and data management skills requires substantial investment. Leo, who teaches courses on big data for social science research, uses a significant portion of class time to teach students how to obtain social media data from online APIs. Even with such targeted learning opportunities, collecting data is a struggle for many social scientists, including experienced researchers. Shawn, who already had programming skills and experience with other web-based data, relates his data collection troubles:

> I had not collected Tweets before. I had collected other types of data [...] When I went to go collect this [Twitter] data, I thought, well there are some commercial tools available, there are some open source tools available, there must be some literature on this. There were some huge issues with the commercial tools, they were really expensive and you couldn't get your data out. [...] The open source tools were sort of custom tailored for most of the other folks' research and they were all very, very difficult to use and required a very high degree of technical expertise.
>
> [...] Then I kind of got mired down into this tar pit of data collection, and so that led to my dissertation. My dissertation isn't about political participation at all, my dissertation is about how we collect this data for research purposes…
>
> – *Shawn*

Contrary to expectations, available off-the-shelf tools for collecting data did not meet his needs. For Shawn, the difficulty of getting data from social media APIs turned out to be such a significant unsolved technical and methodological challenge that it has become a focus of his dissertation research (Driscoll & Walker, 2014). Laura had a similar story. In her research, she carefully thought through the data collection issues pertaining to her research questions and built an elaborate social media data collection infrastructure that met her particular research needs:

> We spent about a year thinking about kind of the fundamental core questions we had for the project and with those questions in mind spent the next year designing a data collection system, and then have spent the past three years or plus working with that data. We have really rich data starting in 2009, of Twitter data, pretty much continuous collection since then […] so we continue to collect data, so we had built and started that, and then that kind of just runs and now we have all these other questions that we're interested in exploring.
>
> *– Laura*

Now that Laura has been using the system continuously to collect data for several years, she has amassed a very large amount of data targeted towards the specific type of research she does. Issues she struggles with include rate limiting, selecting filtering keywords, and obtaining a statistically useful sample. As with Shawn, the design and creation of a research-oriented data collection infrastructure was a significant challenge, and both of them see it as a major accomplishment and contribution. The laptop that one participant used for collecting Twitter data is in Figure 3.2.

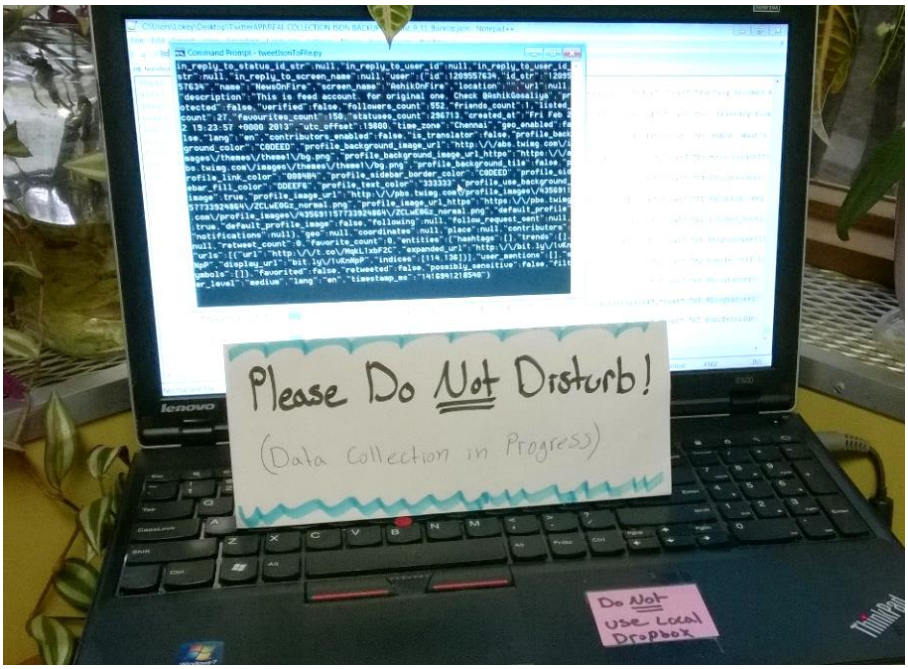


**Figure 3.2: A participant's setup for collecting social media data. A Python script running on the laptop saves raw tweets in a JSON file and prints them to the screen.**

While Twitter makes a great deal of data open to the public, most other social media sites are not so easy to work with. Leo maintains a large network of collaborators, many of whom can provide access to data from less public social media sites like LinkedIn, Facebook, and Google+; some of Leo's collaborators are employees of social media companies who have helped to broker data sharing agreements. However, these connections can be unreliable: when his colleague at LinkedIn left the company, he lost access to data he needed to continue that line of research. Structural barriers created when data is locked within organizations make it difficult for researchers to share their data and methods openly with the broader research community (Huberman, 2012), and limit broad participation in social media research (boyd & Crawford, 2012). Alan also sees inequality of access as a problem for the research community:

> The big problem is that now, data is not really publicly accessible to everybody. It's owned and purchased and monopolized by some big corporations and in order to

> do research, people need to buy it. So we see, it's such a concentration of the access into few hands.
>
> – *Alan*

#### 3.4.1.2 Quality and Completeness

The way that social media data is collected can have a major impact on data quality, which is an important issue for social scientists. Data quality can be affected by at least two sources during data collection: rate limiting and dropped connections:

> I would say the most frustrating thing about writing this paper was probably the incomplete nature of our dataset. The fact that for one, we are dealing with the normal rate limiting. We were dealing with some dropped data collection at times which creates methodological problems.
>
> – *Eric*

Dropped connections and interruptions can create gaps in the data stream, which complicates analysis. Because Twitter limits the rate at which their APIs deliver data, researchers usually obtain only a tiny fraction of the complete Twitter activity that matches their queries, and the statistical quality of the sample provided by Twitter is unclear. Even if the data from social media were not rate limited and could be gathered without gaps, researchers have raised critical questions about whether social media data can produce sociologically valid conclusions (boyd & Crawford, 2012), given that social media's user base is not statistically representative of the broader human population. Anna explained that research in sociology is traditionally done with surveys over nationally-representative samples, and one focus of her research is how to work with big social data and get results that are comparable to these traditionally controlled samples.

However, the biased representation provided by social media platforms may not always be a flaw. Jamie explains that in her study predicting public opinion, this bias may actually be desirable:

> There's another thing about Twitter being not representative of public opinion. [...] Another paper I read said that's why it was bad and they chose something else instead. But for my case, it's actually perfect because I don't necessarily want public opinion. I want people who might be influential […] That's why Twitter is good and unique in that sense because it's capturing that, at least according to the Pew Research study, [influencers are] the primary people who are engaged with it.
>
> – *Jamie*

Various data quality issues related to rate limiting and connection reliability, as well as the underlying representativeness of the sample provided by Twitter, are a constant concern for researchers working with social media data.

### 3.4.2 Structure and Quality of Social Media Datasets

Once data has been collected, researchers face a host of technical and methodological challenges that arise from the way that social media datasets are composed and organized. Major issues include dataset size, information structure, data quality, context-dependence, and volatility. Informants and the broader research community are actively exploring tools, techniques, and methods for tackling these difficulties.

#### 3.4.2.1 Volatility and Volume

Social media data is thought to provide a new, broad perspective on human activity because of its comprehensiveness and scope; Jamie said that this is the main reason she decided to work with

tweets in her research, instead of other online media. Unfortunately, the broad view of online social behavior provided by social media leads to several problems.

Volatility, or rapid change over time, is one of these problems. Social media data streams change over time, often rapidly, as events unfold around the world. Conversation topics rise and fall, and participants come and go constantly. Not only does the conversation move and shift, but the actual identities of users are a moving target, as social media user accounts constantly update and change their self-presentation on the platform. Karen talks about "fluidity" in social media datasets:

> People are intentionally fluid about the way that they describe themselves and the way that they position themselves on a platform like Twitter and then there's this real spectrum of positionality that happens and representation that happens.
>
> – *Karen*

When the character and contents of the dataset, as well as the context that gives it meaning, change radically from one time period to the next, it becomes complicated to make comparisons over time.

Unfiltered, social media is also a high *volume* data source, and the sizes of the social media datasets reported in the literature vary from thousands to billions of messages. Most researchers in this study were somewhere in the middle, typically working with modest datasets of several million tweets at the most. My informants rarely emphasized the size of the dataset as a major challenge. Eric, who said his 10 million tweet dataset "wasn't huge," even mentions sharing the dataset (a few gigabytes) on a flash drive early in their project. Participants' datasets were not so large that they were forced to use "Big Data" tools, like NoSQL databases, Hadoop/MapReduce, and other distributed and high performance computing technologies (although some were experimenting with NoSQL and cloud-based VMs for larger capacity data collection and analysis functions).

Most relied on traditional data storage and manipulation technology, such as flat files (JSON or CSV), relational database management systems, and Python scripts for the bulk of their work.

But the impact of dataset size is not clear-cut. It should be noted that even these moderately sized datasets can be challenging to some of the tools that participants are accustomed to using, such as Excel, Google Drive (spreadsheets) or Tableau:

> Drive couldn't handle the amount of data that we're using. It would have crashed. Tableau is supposed to be better. My sense is that it would still potentially have some issues with just the sheer volume of data we were talking about.
>
> – *Eric*

In this case, Eric's group uses Google Drive to store spreadsheets of social media data, and to do some simple visualizations of the results. They also use Tableau to create visualizations. Eric typically uses Python scripts to aggregate or extract samples of the data before loading it into these tools because they couldn't cope with the full dataset. His group then discusses these visualizations during meetings. Eric said that the aggregation scripts were sometimes so time-consuming that he ran them offline, rather than interactively during meetings: "Some of [the scripts] definitely weren't the most efficient, but certainly none of them were over 24 hours; occasionally, I'd let some run overnight." The delay of running new analyses offline prevented his group from exploring their data more efficiently during meetings.

#### 3.4.2.2 Information Structure and Context

Although social media datasets can vary widely in terms of size, they have certain commonalities in structure and content that present significant challenges for social science researchers.

Twitter and other social media APIs typically deliver data that includes, along with the raw text of each message, a variety of structured metadata, such as account information, location, language, and time. This information can be important: Shawn reported that this metadata was crucial to his work, and was frustrated when tools did not preserve all of the metadata that he needed.

> Since the three of us were basically trying to study a similar dataset, but approaching it from three very different perspectives, preservation of all of the metadata was actually extremely important and most of the tools were basically like, well which two pieces are important to you and I'll keep those two pieces and then they throw everything out.
>
> – *Shawn*

The metadata attached to posts lets researchers look at the data through different lenses (e.g. as events unfolding over time, as local discussions, as information propagation), but preserving the correct metadata and tracking what metadata means to different stakeholders is challenging. In data-intensive interdisciplinary research, differing interpretation and use of metadata can produce friction (Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011).

While the data structure (and metadata) provided by the social media platform can be useful in some cases, answering social science questions sometimes requires other kinds of information about message content and user accounts. Social media data providers have structured the data to make certain information accessible, while other useful information may be implicit, perhaps embedded in the unstructured text-based messages, or missing altogether. Anna describes social media data as large and sparse, meaning it has few "indicators," or attributes of people and messages that sociologists would like to know, such as age, race, and gender (Sloan et al., 2013).

In a project with Google+ data, Leo says that "the most difficult thing is understanding what can be extracted from the data with a certain degree of confidence."

Laura and Karen both spoke about the inappropriateness of social media data structure for the kinds of research questions they work on. Laura says that Twitter data is "structured in some ways and unstructured in others," while Karen says "Twitter is a deceptive thing because it seems like it's structured data, but it's actually not structured for the kinds of questions that I'm interested in." She elaborates:

> [If somebody on Twitter] is a journalist and doesn't identify themselves as that [...] or, it's happened, a PR consultant who is trained as a journalist or something like that, and their profile says they are a soccer mom, it's a lot of detective work to figure out that their information practices and uptake actually are drawing on their trained media skills or previous media experience. For that kind of more sociological stuff, Twitter is actually not very structured. [...] It's deceptive because there's over 200 different kinds of metadata around Twitter. You can be deceived into thinking that those categories are meaningful. [laughter] They're not very helpful... when you try to link them back to something like activity that's happening on the ground.
>
> *– Karen*

Information that Karen needs to answer sociological research questions is missing from her Twitter data. In this case, the data contains limited details about user accounts, and a great deal of time is required to manually reconstruct more detailed information based on the data that is provided.

As Karen says, when comparing social media data to a specific series of events "on the ground," the structural information deficiencies of the data become a major challenge. For example, Eric laments how little of the data delivered from Twitter includes precise geographic coordinates: "Eventually, we kind of started looking at geographic locations, but there is so little geographic data there, that's not really that applicable." Lack of geolocation in the data also affected Jamie's research project. Location in Twitter is such a major challenge that automatically inferring geolocation for tweets is an active research area in its own right (Rahimi, Vu, Cohn, & Baldwin, 2015). The structure of social media data is based on a mix of business, privacy, and technical concerns. For example, Twitter does not enable geo-tagging of posts by default to protect user privacy ("FAQs about adding location to your Tweets," 2015), which may explain why few posts are geotagged.

Because needed information about social media activity is often implicit or unavailable in social media data, informants stressed the importance of context when interpreting observations about social media data. Karen explains:

> To understand what a tweet says you have to either look at all the tweets that are related to it, or a lot of other tweets that are related to it usually. […] You often have to look at what the link is. At least the Tweets in the events that I'm looking at, they're information dense, or communication dense because they might be sharing photos of something like a search and rescue dog or something like that. It tends to be very hard to determine. You can't look at the individual Tweet level. It's actually pretty hard to render meaning that way.
>
> – *Karen*

Here, Karen reveals the lengths that a researcher must go to in order to understand the meaning behind any particular tweet; a great deal of additional information must typically be gathered and considered, beyond the tweet in question. Many researchers bring in additional data streams to supplement their social media datasets, providing some of this much-needed context. Thus, while the social media datasets my participants work with are not extremely large on disk, the implicit context around the dataset makes it large in another sense, and time consuming to work with.

## 3.4.3 The Analysis Process

This section provides an overview of two major issues in analyzing social media data: determining relevance and exploring the dataset. It then discusses some specific technical issues for software that researchers use to work with social media data.

### 3.4.3.1 Determining Relevance

A major obstacle that informants brought up is determining which portion of a social media dataset is relevant to the research question at hand. Relevance is a crucial issue because of the size, volatility, and quality issues associated with social media datasets. The queries that can be issued to obtain data from the Twitter API may include filtering based on keywords, hashtags, or user accounts, but these have limited expressive power and tend to match a lot of irrelevant content. How to define relevance may be unknown until after the dataset is already collected, making it difficult to specify precise keywords in advance, so informants report that they prefer to use inclusive filtering criteria up front, and then spend time refining their dataset after it has been obtained. Jamie calls this approach "Collect now, Filter later." Karen says that refining the dataset post-collection is time consuming:

> It takes a lot of time to actually get to know the datasets. I have spent ... I started looking at this dataset in January and it's June. So, about 6 months, I feel like I know a good portion of this dataset actually is unrelated to the event. Things as simple as finding things that are even related takes a lot of time in the datasets.
>
> *– Karen*

Karen obtained her dataset using several hashtags and keywords that she knew were relevant to the event, but these filters also matched a large number of tweets about other topics. She reports spending a large amount of time to establish the "beginning and end of the information space" within the dataset. Shawn, who did research with tweets related to the Occupy movement, adds that the definition of relevance depends on the researcher's orientation:

> So you might collect this large dataset [...]. But then you don't ever want to use [all of] that data, it's usually problematic to assume that a hashtag or a page or a keyword is all relevant to your research question. And relevancy is subjective to your research question right? Because I might take the same dataset, ask a different question and then I have the different concept of relevancy. Like in our Occupy dataset, for example, we have the use of "I'm hungry, today I want to occupy Chipotle." Is that related to Occupy or not?
>
> *– Shawn*

Researchers report arguing extensively over difficult and ambiguous decisions like this one as they explore and come to better understand their datasets and their research questions.

#### 3.4.3.2 Exploration

The process of establishing relevance in the dataset goes hand in hand with exploratory data analysis. Most of the researchers in my study were conducting open-ended, exploratory research projects where they refine their goals and research questions as they analyze the data. In these projects, it is important to understand what is going on in the dataset, how the pieces fit together, and how they change over time. Because large social media datasets are complex, dynamic, and multi-faceted, exploring them is challenging. Eric talks about coming to terms with a new dataset:

> Initially, yeah kind of just knowing how to approach it and how to tackle it, how to break it up into different chunks wasn't necessarily clear. Especially since we didn't know what we were looking for. It was difficult to know how to break that data up into something that was more manageable.
>
> – *Eric*

In Eric's project, the group used a visual cluster analysis of hashtags in Gephi to provide starting points for breaking down the dataset into manageable chunks. Visual and statistical techniques can provide points of entry, after which researchers often resort to reading raw tweets:

> It's hard to browse, right? [...] We have all these strategies of like, let's look at time series data, let's look at, read through some of the tweets…
>
> – *Laura*

#### 3.4.3.3 Tools and Software

Researchers talked about using a combination of many tools and software to work with their data, ranging from short, one-off scripts (often in Python or R) to complex software like Microsoft Excel

or Tableau. No single tool can meet all needs. Shawn, whose research concerns the computational and information infrastructure underlying social media research, argues that researchers depend on having an ecology of tools they can use to process data and answer their research questions:

> What [researchers] really want is a set of tools that will help them collect the data, explore the data, select the slice that they want, and then export that into a format that they can use the tools and methods that they already want to apply.
>
> – *Shawn*

While working with multiple tools may be unavoidable, moving between tools requires transforming data and dealing with compatibility issues. One interviewee sketched out the data analysis workflow in for their research project in Figure 3.3, illustrating numerous transitions through various data formats, scripts, and analysis steps.

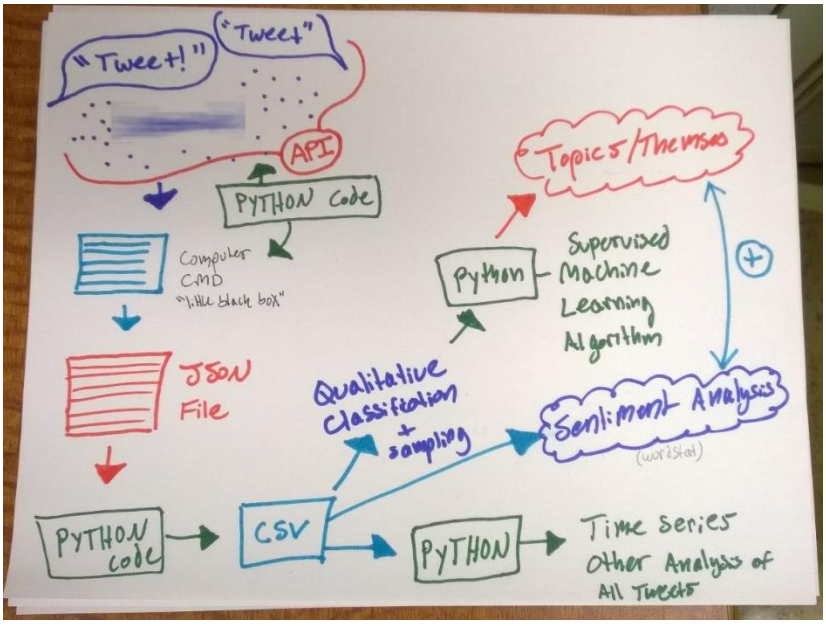


**Figure 3.3: An interviewee's sketch of social media data analysis workflow. The data is collected and processed by Python scripts, transformed between JSON and CSV, qualitatively coded, computationally modeled, and visualized using time series plots.**

Transforming social media data from one format to another to satisfy various pieces of software and analytical techniques is a time consuming and error prone process. Commonly, researchers convert some data to CSV formatted files, which can be opened in spreadsheet software like Excel or Google Spreadsheets; this makes manual inspection, reading, and qualitative coding much easier. Jamie reported encountering major roadblocks with JSON to CSV conversion because of formatting errors in the JSON file. Researchers also had problems viewing their data in Excel, since it does not always work correctly with character encodings used by Twitter, and sometimes reformats data in CSV files when it saves them. Alan tried to code several thousand posts containing Chinese characters in Excel:

> I was able to open the CSV file with Excel, but the problem it has is that the text part just got messed up, so... weird signs. I was not able to read the text part, but for other parts, for example, the user name, the time that the tweets were posted, they're all there. I was able to use Excel, and then I'd create a new column just for codes, and then I set up Excel on my right and I opened that file with R on my left and I manually coded, one by one.
>
> *– Alan*

Alan found that R was able to display Chinese characters correctly. Excel's limitations forced Alan to develop an inconvenient, if resourceful, workaround using R and Excel side-by-side.

While transforming the data to CSV format makes it easy to work with in small batches, with larger datasets some form of database is often used. Different styles of databases can introduce further software compatibility issues. In the example below, Eric explains how switching from a

relational SQL database to MongoDB, a popular NoSQL data store, meant they could not directly visualize the data with Tableau:

> I guess the other issue with using something like Tableau is that while we are originally working in a relational SQL database, when we moved over to the server, we moved to a non-relational database which would have been [difficult] for Tableau to handle.
>
> – *Eric*

Instead, Eric would write scripts that transformed the data from MongoDB to CSV files which could easily be imported into Tableau for visualization. Incompatibility between software tools and the necessary data transformation workarounds lead to friction for researchers, impeding the data analysis process.

### 3.4.4 Analytical Techniques

In this section, I will discuss several analytical techniques used by these researchers. Some of the most essential techniques are simple text searches, reading the data, and viewing simple visualizations; researchers combine these techniques to explore the data, check their intuitions, and locate areas of special interest within the dataset. Some researchers also use various computational analysis techniques, such as clustering and topic modeling, to help reveal structure in the data. Finally, being focused on the text-based content of their data, most of my interviewees reported doing some form of qualitative coding.

Researchers studying social media data come from a variety of disciplinary backgrounds and perspectives, and bring with them different methods that they apply or adapt to study social media

data. Even when the work that researchers are doing might look similar, they might be drawing on subtly different methodological foundations. For example, though many researchers might be doing qualitative coding, different interviewees talked about their analysis using different methodological vocabulary, such as content analysis, ethnography, and grounded theory. These reflect different goals and expectations for their analytical work. These disciplinary differences must be kept in mind to avoid confusing terms that may not always mean the same thing.

#### 3.4.4.1 Search, Reading, and Visualization

Because of the complexity and volatility of social media datasets, many researchers use text-based searches to quickly and easily slice their datasets into focused samples. Simply reading tweets provides a crucial opportunity to qualitatively interpret the data and check assumptions about possible findings. Visualizations such as time-series plots are used to quickly summarize and compare sections of data.

Text-based search is a common technique for analytically restructuring large social media datasets. Search provides a way to rapidly, if imprecisely, identify examples about a given topic, rendering the primarily text-based content of the data accessible for analysis. As Karen explains below, data matching a search query may be quickly scanned in an exploratory process:

> [In interview transcripts], I did a bunch of coding around "food," so I pulled a bunch of tweets around that. I do a lot of text searches, text string searches in SQL [...] Every time I do an interview I go back and search for words related to foods. Every conceivable word that I could find that was related to food. All the tweets that I could find I would just sit there and read through them.
>
> *– Karen*

In the above example, Karen discovered an important topic ("food") from interviews she had conducted previously. She used search queries followed by reading the matching results to relate this interview finding to her social media dataset, and reports that her attention often shifted dynamically back and forth between interviews, searching the Twitter data, and reading tweets. Eric talks about reading tweets while using charts and visualizations during group meetings:

> Generally, we would have that conversation about what we thought was kind of important and what we thought was interesting about those particular graphs and I would generally start and talk about what I had found and then other people would bring up other ideas or comments on what I had come up with.
>
> Once we had identified these other interesting trends or points of interest, sometimes we'd use meetings to go back in and pull these tweets out. We would just have a bunch of laptops sitting around so we could just go into the database and run a query and list a whole bunch and then see what was actually going on.
>
> – *Eric*

Eric says that in group meetings the project members review visualizations Eric has generated (in Google Spreadsheets usually), and check their interpretations by looking up example tweets based on what they see in the visualizations.

Karen emphasizes reading the data as part of "learning the data and what the data can say". She says that "[I do that by] reading a lot of Tweets. We spent 6 months reading this dataset and exploring and iterating on it." As Karen indicates, reading is a time-consuming way to explore the data. Some researchers approach their data with exploratory statistical analysis, which can often be performed more quickly:

> We didn't really have any concrete goals when we started out. We started essentially doing exploratory analysis on the dataset looking for interesting trends. I was most involved and kind of doing some basic statistics on it, looking at top hashtags [...], trying to essentially find patterns within the dataset that we could then look at.
>
> *– Eric*

Here, Eric talks about looking at top accounts and tweets according to various statistical criteria calculated from metadata. Participants often referred to basic statistics such as these, which they used to narrow down the scope of the dataset to a manageable size for manual exploration.

When quantitatively analyzing social media data, as with visualizations or statistics, researchers need to be able to work with the data at multiple levels of aggregation or multiple units of analysis. The volatility and fluidity of social media datasets requires that researchers zoom in and out to different scales. For example, researchers may want to understand what is changing in their data over both weeks and minutes. Laura explains how the datasets change at different rates:

> A lot of times we have like a consistent population where the network is changing at a slow scale but information is diffusing […] at a rapid scale. Like we have these different time scales of data which is hard to think about all those things at once.
>
> *– Laura*

In addition to aggregating multiple levels of time, researchers discussed analyzing geographic location at multiple levels, and various groupings and hierarchies of

accounts and messages. Many tools which researchers use have limitations on working at multiple scales and levels of granularity:

> There were things that [Google Spreadsheets] could not do but that usually just had to do with the resolution of the visualization that it could produce. Sometimes, if we wanted to see something that had really fine granularity especially in terms of time, it could be difficult to actually produce anything meaningful in [Google Spreadsheets] whereas Tableau would be able to graph that pretty well.
>
> – *Eric*

#### 3.4.4.2 Computational Techniques

In addition to the basic analytical techniques above, researchers report using a variety of computational techniques, such as clustering, topic modeling, dimensionality reduction, and semantic networks. Most of these techniques aim to reveal "natural" structures within the data, making them well-suited for exploratory data analysis. Several informants mentioned looking at relationships between words in the tweets. Below, Eric explains how these networks were derived from a tweet dataset in the early phases of a project:

> It was exploratory, but we essentially just created a network of the co-occurring hashtags where we described the strength, the tie or the edge, what's the number of times that the hashtag occurred together and then grouped based on that.
>
> – *Eric*

The resulting "semantic network" of words or hashtags was then imported into Gephi for visualization, where it helped the group begin to wrap their heads around the dataset:

> You have a massive dataset, and so [visualizing the hashtag networks] allowed us to kind of break it down into more manageable chunks and gave us smaller sections and smaller portions of data to look at. Before that when we had just started, we had no idea what we were looking for.
>
> *– Eric*

Computationally analyzing and visualizing semantic networks provided an efficient way to reveal patterns in the tweet content, suggesting some starting points for the deeper analysis that followed. Laura used this exploratory visual approach, but she also discussed analyzing social media data using topic modeling (i.e. Latent Dirichlet Allocation): "This is one way we try to say 'what are people talking about?' Let's try to fit these topic models that show us broad themes…" Laura has also used clustering and dimensionality-reduction techniques, such as multi-dimensional scaling, to find groupings of social media accounts. While these computational techniques are not as direct as simply reading tweets, they still let the data "speak for itself" in a sense, supporting exploration of the data. Laura said that she combines many different techniques to support exploratory analysis:

> I sort of have this little tool kit of common things that I do, like network diagrams, multi-dimensional scaling, time series, kind of standard tools and in combination, you know, they tend to give a limited picture, but a pretty good picture of what's going on.
>
> *– Laura*

While every analytical technique has strengths and weaknesses, using a combination of multiple techniques helps Laura see the data from different angles, and compensates for the limitations of any single approach.

#### 3.4.4.3 Qualitative Coding

Qualitative coding was one of the most crucial analytical techniques reported by interviewees. Qualitative coding is a general process of sorting raw data into a set of categories or "codes," or tagging selections of data with codes. As reviewed in Chapter 2, coding is part of many forms of qualitative research, but can be conducted in different ways, for different goals, and at different stages. Sometimes, the researcher approaches the data with a theory developed beforehand and imposes it upon the data. In this scenario, qualitative coding is a "closed-ended" kind of analysis. Alan's approach was explicitly closed-ended; he developed a model of the phenomenon he was studying in advance, and then applied it to the social media data for validation:

> I came with a research question. I came with my own theoretic framework. I didn't just let themes or patterns emerge themselves, I had my own theoretic categories […] The way that I came up with this model is I read a similar published paper [...] Their model is similar, not exactly the same, but similar. This model is developed upon that model.
>
> *– Alan*

Subsequent to coding the data using his scheme of categories, he counted the occurrences of each code and used these quantities to test his theory. However, sometimes the role of coding can be blurry or dynamic. Eric discusses a project where, as in Alan's case, the coding scheme was initially created from an article on a related topic published by other researchers:

> The coding scheme was [...] a combination ... It's in part based on a previous paper. They had a similar scheme that they used for essentially similar goals. We started there.

However, as coding continued, Eric says that the scheme was modified and reorganized to better reflect the data being coded, as well as the evolving goals of the current research project:

> The coding scheme we then tweaked [...] The rest of the coding scheme came from trying to code tweets and then reflecting on whether their codes seem to fit what we are trying to look for. Whether the coding scheme really fit the data or not. Whether we were actually accurately describing the data and going back and changing the codes or changing the definitions and adding more codes and subtracting codes.
>
> *– Eric*

In this second stage, Eric's research group began to draw ideas from the data, to evolve the system of categories they were using. Here, the coding becomes open-ended. Other informants reported taking this approach from the beginning. Karen, who describes her methodology as "grounded," says that she applies an open coding process where categories are created by inspecting the data, without starting from some existing system of categories. She elaborates on her coding process:

> I started open coding on the top 100 most tweeted, top 100 most re-tweeted [...] They're not completely open codes, these come out of the research questions that I have. I want to understand where people are getting information from and it's important for my work. For my work it's really important to understand who the

> people are that are sharing information, where are they coming from. A lot of it is driven by those, where categories came from.
>
> – *Karen*

While the categories are still shaped by researcher expectations and research questions to a degree, here the goal of coding the data is exploratory, intended to reveal structure contained in the data. Karen stressed the importance of coding for exploration of social media datasets: "My way of exploring is to start coding because it's a way for me to get my head around it." As we see here, coding can be used both to explore the data in an open-ended manner, and to categorize the data using categories that the researchers bring to the data. In exploratory research, qualitative researchers may use a combination of both approaches at different stages, progressing towards refined, structured coding as the research questions become clear and specific.

## 3.4.5 Socio-Organizational Challenges

Interviewees mentioned several psychological, social, and structural challenges related to doing research with social media data, but not directly derived from the data itself. These included fear and anxiety experienced while trying to learn data collection and manipulation skills, concerns about privacy and ethics, and challenges with obtaining funding and access to technology.

### 3.4.5.1 Fear and Anxiety

Researchers getting started with social media data also face formidable social and emotional challenges. Learning to program can be a terrifying undertaking; many beginners experience extreme fear and anxiety (Rogerson & Scott, 2010). Jamie, who, at the beginning of my interview study, had already tried to learn to program several times and given up, reported that she had been learning to collect social media data for months. She said that her biggest obstacle was fear:

> I think it's just fear, that it was so completely unknown. I'm not even a Twitter user. [...] It's the same feeling of like paralysis and helplessness [...] of just completely not knowing what to do or where to start and it was this alien thing, and that is a huge barrier.
>
> – *Jamie*

Although Jamie did come to social media data with prior programming experience in R, and experience building computational models, she describes this previous experience as "really scary," and explained that having a bad experience with programming in the past did not make learning it again much easier. Jamie eventually succeeded in overcoming the fear barrier. After another researcher shared a sample data collection script with her, she struggled for some time to get it working before eventually solving the problem. Emotionally, this success had a big impact: "Getting the collection to work was like a huge, huge thing." Jamie also combatted her fear by joining a community. She met with other learners and attended classes and workshops:

> Realizing that that discomfort that I'm experiencing, that fear is totally normal. And not only that, there was a whole bunch of people just like me who are going through it right now, and we're all sort of helping each other trying to find those resources. Having a community around that was really important.
>
> – *Jamie*

While there are many solutions to the fear problem, programs that help beginners connect with each other can be an effective way to overcome some of the hardest and most personal hurdles encountered with social media data, as well as with data science and programming more generally.

#### 3.4.5.2 Privacy and Ethics

The collection and analysis of data from social media sites is fraught with ethical dilemmas. Based on a case study analysis of a 2008 release of Facebook data in 2008, Zimmer has argued that ethical research with social media requires reconsidering traditional conceptions of consent, expectations of privacy, and strategies for anonymization (Zimmer, 2010). Moreno et al. elaborate on these concerns, and develop specific recommendations to help researchers and Institutional Review Boards develop ethical research protocols (Moreno, Goniu, Moreno, & Diekema, 2013). Lafferty and Manca's review of social media research literature also discusses ethical considerations (Lafferty & Manca, 2015).

Several interviewees raised concerns about privacy and ethics, issues which they say have not yet been resolved by the research community. In research involving human subjects, Institutional Review Boards (IRBs) monitor the conduct of researchers to protect subjects, minimize risk of harm, and ensure fair, non-exploitive procedures. However, there is ambiguity and a lack of agreement about how to handle research conducted with social media data. This area of research being fairly new, IRBs have not yet established standard policies for how it should be conducted under the guidelines for human subjects research, or even whether it qualifies as human subjects research in the first place.

In part because of the lack of clear guidance from IRBs, researchers report struggling with privacy and ethics issues on their own. Some feel that data like Twitter posts, which can be obtained by anybody using a public API, are public data, while others may believe the Twitter users have a reasonable expectation of privacy. The question of what measures researchers should take to protect social media users from harm is unresolved. Informants reported engaging in practices such as anonymizing tweets by changing usernames, rewording messages so that they are not easily

found via search engines, and asking for permission directly from social media users to include quotes in publications. Researchers also pointed out that some social media users they have spoken with prefer to be acknowledged for their contributions, rather than hidden behind anonymization.

Concerns about privacy and ethics are intertwined with technical decisions that researchers make. In the earliest phase of their project, Eric's group shared portions of their data by copying it onto personal machines. Eric comments on how his group's decision to transition to storing their social media data on a centralized, access-controlled server made him more confident that they could protect users' privacy:

> I think having the server is good for two reasons. One, it means that the data is accessible to all of the collaborators. It's also good from a privacy standpoint in terms of where we're hosting our data. I don't think that we should necessarily have it spread all over individual machines.
>
> – *Eric*

Since Eric's project involved a large number of collaborators who would come and go over time, being able to permit and revoke access to the data allowed them to better enforce privacy policies:

> We could definitely do a better job of making that data more secure in terms of taking people off... Removing access for people who are no longer collaborators, restricting access and giving certain people different write permissions which is just something we haven't done because [...] we don't really have the time to do so […]

> Since it's been changing every quarter, because it is running as a [quarterly research group], we constantly have new people coming in who need to work with it.
>
> – *Eric*

Beyond privacy, Shawn also raised the complex issue of ethical representation of participants in research. He offered a critical perspective, arguing that many researchers studying social media data do not take seriously the human aspects of the data that they study, and their responsibility to treat these humans respectfully:

> They're missing the person component and the community component. They're not interacting with that community, they're not asking that community for permission, they're not asking that community what does this space actually mean? [...] people aren't just a series of numbers. You have to understand what these numbers represent and then you have to [...] take in the context and all that into consideration.
>
> – *Shawn*

Here, Shawn is suggesting that a researcher should try and see social media data through the eyes of social media users, and respect the perspective of social media users when they represent users in their research. For Shawn, whose viewpoint comes from his background in social science theory and public policy, ethics is a pressing problem for the social data research community. He argues that this discussion should be broad and inclusive:

> I don't think we've taken the discussion to all the places where the discussion has to happen. We've done a series of workshops at conferences […] But that's very

> sort of insular, within-community conversations. But the question is how do we have these conversations across fields, across countries, across seniority, right?
>
> – *Shawn*

As the community rushes to explore the area of big social data research, events like the 2008 release of unsuccessfully anonymized Facebook data from 1,700 students ("Tastes, Ties, and Time: Facebook data release," 2008), and the controversial Facebook emotional contagion study (Kramer, Guillory, & Hancock, 2014), catalyze discussion and shape public perception and institutional policies.

#### 3.4.5.3 Funding and Technology Access

Researchers also reported having problems getting funding for the technology resources they need. While social media data collection and analysis might be conducted on a personal computer in the initial stages, setting up a more resilient and reliable data collection system and scaling to larger datasets requires computational infrastructure that can become expensive. Shawn describes his experience trying to get funding for cloud computing resources to collect social media data:

> I wrote some duct tape and bailing wire, sort of bash scripts to collect data on a free instance on Amazon on AWS. Then we asked the school for resources and they're like we really don't have any, so that forced us to find a faculty member and that's how [our faculty supervisor] came on, and we asked if he wanted to help us and play kind of offense and defense with administration and help us get some resources.
>
> – *Shawn*

Eventually, Shawn's group was able to obtain grant funding from the National Science Foundation to continue developing tools and methods to collect and analyze social media data. Shawn was able to succeed because he found a faculty member who agreed to sponsor his work, but some researchers may not easily find this type of support. Jamie reports that she is the only student in her program who is working with online data, and that her adviser is unfamiliar with using social media data for research. She describes encountering problems getting access to data analysis software that would enable her work:

> It's sort of a constraint for me. I actually really want to be working on [analyzing the data] but we don't have the technology or the funding to have enough computers to be able to do the kinds of things that I want to do. [...] [The data analysis software] was really expensive. I mean, I might have tried to bite the bullet, but [my adviser] was kind enough to buy it for the lab. [...] I need to have that so I can start working on it and getting it more accurate.
>
> – *Jamie*

The software that Jamie needs is proprietary and expensive, and the licensing scheme makes it difficult to get regular access to the computers where it is installed. While the struggle to obtain funding for research is not new, perhaps it is the *interdisciplinarity* of social media research in the social sciences that makes funding especially hard to find (Metzger & Zare, 1999). Jamie and Shawn both work in fields where the methods and techniques they are applying were unfamiliar. Both researchers collaborated with individuals across departments, but making the case for funding technology resources in their home departments required some extra effort. With researchers across many fields becoming aware of the value and cost structures of using social media data,

these barriers should erode. Jamie's emphasis on community suggests that researchers in this space may find it helpful to organize inter-departmental groups and support structures, to share strategies and ideas around funding and technical infrastructure.

## 3.5 Implications for Design

This section highlights key implications for design, based on the findings discussed above. I focus on challenges and opportunities related to data collection, exploratory analysis, qualitative coding, and continuity throughout the data analysis process.

While tools to help researchers collect data are improving, data collection remains a major barrier for many researchers. Better data collection tools may help, but the underlying APIs provided by social media companies are not designed to deliver high quality samples suitable for scientific research. Social scientists are forced to work around rate limits, inadequate filtering options, and opaque sampling methods. The validity and reproducibility of their work can be damaged by access issues (boyd & Crawford, 2012; Huberman, 2012). While better data collection technology could lower barriers, social media **companies must commit to opening their data to researchers** for the technical and methodological challenges of data collection to be solved.

Because of the complexity and diversity of social media datasets, researchers reported conducting extensive exploratory analysis, but with the technology researchers have available, exploration takes a long time. Researchers combine search, statistics, visualizations, qualitative coding, and computational analytics to explore their data, and these techniques involve using a variety of tools and software (such as Excel, Google Spreadsheets, Tableau, Gephi, and sometimes qualitative coding software). Unfortunately, most of these tools and techniques were not designed or developed with social media datasets in mind, and often adapting the software to suit the needs of

social scientists is difficult. Social scientists need tools that are specifically designed for exploring social media datasets. Several opportunities are outlined below.

Researchers report spending a lot of time manually looking up missing contextual information. Software could provide specialized support for social media by automatically **gathering and presenting context**. For example, the SRSR system, designed for journalists, shows summary information about user accounts in social media datasets (Diakopoulos, De Choudhury, & Naaman, 2012). Looking up these contextual details automatically would reduce the friction and awkwardness of switching tools, and would allow analysis to flow smoothly.

Tools could also support exploration of social media datasets more effectively by **working with all aspects of social media data** that social scientists care about. Researchers referred to several tools that they used to work with the text content of messages, including qualitative analysis software like QDA Miner and NVivo, as well as spreadsheet software like Microsoft Excel. Meanwhile, researchers are able to analyze many quantitative and categorical metadata dimensions in visualization tools like Tableau and Gephi. However, few tools link these different aspects of social media datasets together, or integrate interactions between all aspects of the data.

Tools designed for analyzing social media data should provide support for **working at multiple scales**, or units of analysis. Flexibility to different units of analysis is a general design principle for exploratory data analysis tools, as "appropriate scales of analysis are not always clear in advance and single optimal solutions are unlikely to exist" (Keim, Kohlhammer, Ellis, & Mansmann, 2010). Because social media datasets are complex, volatile, geolocated, and temporal, analyzing data at multiple levels is especially important. For example, Eric said that visualizations available in tools such as Excel and Google Spreadsheets did not provide flexible aggregation to

zoom in and out of the data, and he had to switch to Tableau or re-aggregate the data in Python to gain this capability.

Another concern that tool designers should keep in mind is the friction that incompatibility and data transformation can create for researchers. Interviewees often referred to transforming data between formats to import it into one tool or another. While it is impossible to provide perfect compatibility and eliminate such transformations entirely, tool designers should do what they can to **minimize data transformation** when switching tools. For example, adhering to standard data formats, providing data ingestion tools that are robust to unexpected input, and supporting modern character encodings are all ways that the cost of software switching can be reduced.

The social scientists interviewed in this study used qualitative coding heavily as both an exploratory and targeted analytical technique. Most current software tools for qualitative coding are oriented primarily towards document-style qualitative data such as interviews and field notes, and are expensive and proprietary. Researchers need **qualitative data analysis software that supports social media data**, which is composed of thousands or millions of short messages with a wealth of structured metadata.

The smooth progression from open-ended, exploratory analysis to formally-structured confirmatory analysis was a general pattern found in the practices of many of my participants. For example, researchers practiced several styles of qualitative coding, ranging from exploratory, open-ended coding to targeted, systematic coding. Software designed well for one end of this spectrum may not work the same way as software designed for the other end, and participants often progress back and forth along the spectrum. This issue is explored further in Chapter 6. This is not limited to qualitative coding; rather, all of the analytical techniques discussed in the previous

sections are combined by researchers, to achieve this progression. Further study is needed to understand how software can support **transitions through multiple phases of analysis**.

## 3.6 Conclusion

My goal in this chapter has been to address the following research question: *What are the goals, barriers, tools, and processes associated with mixed methods online social data research?* I have described an ethnographic study, primarily based on a set of semi-structured interviews with social scientists. My interviews provide rich detail about the data collection and analysis tools and techniques that researchers are using to conduct mixed-methods research with social media data. I have also discussed challenges from constraints in the data itself, as well as organizational and social challenges. My findings point to opportunities for designing data analysis tools for social media data.

One limitation of this study is its focus on members of the social data research community at a specific site. Further research is needed to understand how widespread the practices observed here are, and which findings translate to other settings. While some of the specific techniques that my informants used heavily, such as qualitative coding, may not be shared by researchers who come from a quantitative research background, the social media data itself should ensure there are some commonalities across different research styles and methodologies. In addition, exploratory data analysis is a phase that most researchers go through during their work, and I expect that many of the findings associated with data exploration will be common to most researchers working with social media datasets. Analysis of the literature published by social scientists studying social media data, such as Lafferty and Manca's synthetic literature review (Lafferty & Manca, 2015), could

also be used to map out the different methodological approaches being developed and to better understand of the diversity of this space.

Moreover, many of the practices described in this chapter are not specific to social media data at all, but would be shared by researchers and data analysts in many fields and with many types of data. The emerging practices of researchers studying social media data are not yet well understood, and learning about this important and growing field may provide important lessons for the design of general data analysis tools as well. Further visual analytics research on the practices of data analysts in specific domains, such as intelligence analysis (Kang & Stasko, 2011), building design (Tory & Staub-French, 2008), automotive engineering (Sedlmair et al., 2011), and enterprise data analytics (Kandel et al., 2012), should be considered as a whole, to look for commonalities and build new theoretical understanding that can inform the design of data analysis tools.

# Chapter 4: Qualitative Labeling with Machine Learning

Social scientists who wish to use qualitative or mixed methods approaches to study large social media and online communication datasets face a methodological dilemma. Coding, the bread and butter of qualitative research, is an expensive and time-consuming process that traditionally requires a great deal of manual labor and human interpretation. One solution is to sample and code small portions of larger datasets. This can be a useful exploratory technique, and was used by some interviewees in Chapter 3. However, large social media and online communication datasets can yield not only exploratory observations, but also observations of broader patterns and dynamics. Finding larger patterns requires considering larger quantities of data using statistics or computational analysis. As several social scientists working with online social data have demonstrated (Goggins, Mascaro, & Valetto, 2013; Rosé et al., 2008; Starbird et al., 2015), combining qualitative and quantitative techniques in a mixed methods approach can achieve analysis that is both broad and deep.

However, it can be challenging to effectively transition between qualitative and quantitative perspectives in the same research project. When qualitative categories are developed through manual coding over a small section of data, how can they be connected to findings over larger amounts of data? A crucial requirement is a way to project the categories and theoretical concepts developed through exploratory qualitative work upwards, onto a larger dataset. Machine learning technology, specifically supervised or semi-supervised classification, has the potential to perform this function (Crowston, Liu, & Allen, 2010; Rosé et al., 2008). Using manually coded selections of online social data to train a machine learning system, researchers can apply machine learning to automatically code far more data than they would be able to affordably classify manually.

Challenges with this approach include technical barriers, methodological pitfalls, and complex cost benefit tradeoffs. In this chapter, I consider the following research question: *How can machine learning techniques be applicable for automation in mixed methods research with online communication data?*

To answer this question, my colleagues and I launched a research project studying social and psychological phenomena within a mid-sized online communication dataset. Our project focused on understanding emotion and affect in a chat dataset from an international online collaboration. Geographically distributed collaboration is increasingly common across many work domains, and understanding the expression of emotion in computer-mediated communications is crucial to understanding team interactions and processes. An increasing number of studies in the last twenty years have documented a renewed interest in understanding emotion and affect in the workplace (Ashforth & Humphrey, 1995; Barsade, 2002; Grandey, 2008). Numerous studies have shown that affect and emotion influence performance and interactions in cooperative work environments (Amabile, Barsade, Mueller, & Staw, 2005; Ashforth & Humphrey, 1995; Grandey, 2008; Mentis, Reddy, & Rosson, 2010; Milliken, Bartel, & Kurtzberg, 2003). Further, this "awakening of interest in emotion" (Mentis et al., 2010) can support the design of affect-aware information systems.

Text-based chat within collaborative scientific work provides rich data for understanding affective processes within groups. The increasing volume of text-based communication available for study, combined with a growing awareness of the importance of affect in the workplace, have led to an upsurge in research on affect detection in text, including work in fields as diverse as sentiment analysis, affective computing, linguistics, and psychology, among others (Grandey, 2008).

In this project, our research group used a mixed methods approach. Manual qualitative coding for affect expression in chat logs can yield rich interpretive analysis, but this process does not scale well to larger datasets. Therefore, after doing significant qualitative coding manually to create a grounded framework describing affect in the dataset (Scott et al., 2012), we trained machine learning algorithms on the qualitative coding scheme we had developed, in order to apply it to the full dataset. Research on the automated detection of the overall positive or negative sentiment of long, relatively well-formatted blogs, articles, and online posts has achieved promising results with statistical classification methods based on frequencies of term occurrence, e.g. (Gill, French, Gergle, & Oberlander, 2008; Keshtkar & Inkpen, 2009; Tausczik & Pennebaker, 2010). More recently, these methods have been applied to informal settings, focusing on classifying messages on social network sites, blogs, and discussion forums, which are characterized by irregular grammar and spelling practices, e.g. (Mishne, 2005; Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010). However, classification methods robust to the varied and dynamic context of affect in collaborative and distributed online chat environments remain a challenge (Rosé et al., 2008).

In this chapter, we address the problem of detecting subjective, non-mutually-exclusive labels of affective state (*e.g., joy*, *excitement*, *confusion*, *frustration*, *anger*, or *annoyance*) in workplace chat logs. We describe the design rationale and evaluation of our machine learning system, and contribute a novel approach to automatic affect classification in chat logs that integrates with our interpretation of affect as a dynamic, subjective process. We discuss our novel combination of features for use in classification of affect expression in chat, and present *ALOE*[6], our open source tool for classifying coded chat messages. Our technique was developed for a large dataset of nearly

[6] http://depts.washington.edu/hdsl/tools

500,000 lines of chat collected over four years of an international scientific collaboration, where we were able to successfully identify 13 common types of affect expression from a taxonomy developed via collaborative open coding.

Automated techniques to identify affect in chat messages are a powerful addition to the analytic toolkit of researchers studying large online communication datasets. Based on this case study, we conclude with reflections and lessons learned, focusing on issues of methodological alignment and validity. This work has implications for how researchers might apply machine learning for social science research on social media and online communication data, such as what characteristics of machine learning algorithms are most valuable, and what some of the strengths and weaknesses of these techniques may be. It also suggests priorities in the design of machine learning algorithms and systems, such as transparency and interpretability.

## 4.1 Acknowledgments

Much of the work in this chapter was done collaboratively, with major contributions from Katie Kuksenok, Megan K. Torkildson, Daniel Perry, John J. Robinson, Taylor Jackson Scott, Ona Anicello, Ariana Zukowski, Paul Harris, and Cecilia R. Aragon. Parts of this chapter were previously published as “Statistical Affect Detection in Collaborative Chat” at the 2013 ACM conference on Computer Supported Cooperative Work and Social Computing (Brooks et al., 2013), and then further expanded and edited by the author for inclusion in this dissertation.

## 4.2 Related Work

Previous work includes a diverse set of approaches to affect classification, differing both in the evidence (features) used, as well as the classification method. The variety of ways that the problem has been studied and the many techniques that have been developed make comparison between

these studies challenging. The features used for affect classification include everything from statistics on the frequencies of word sequence occurrences (word counts, n-grams) to linguistically informed features (e.g. part of speech). These features have been combined with countless classification methods, ranging from rule-based methods to probabilistic methods. Depending on characteristics of the data, different configurations can have different levels of success. Furthermore, different granularities of classification – whether to classify affect merely as positive or negative, or a finer set of categories – not only impact success, but make results difficult to compare between studies.

There are many definitions of emotion, affect, sentiment, and related concepts; many definitions are overly restrictive given the wide range of emotional or affective phenomena that may be of interest in understanding cooperative work environments. In this work we draw on Russ's broad definition of *affect* as an inclusive concept spanning emotions and feelings distinct from cognition (Russ, 1993), and more pervasive than the neurophysiological experiences of emotions (Moore & Isen, 1990). We seek to better understand how instances of this broad notion of affect manifest in collaborative text-based communication.

A significant amount of research on affect or sentiment detection in text has focused on lexicon-based approaches, in which pre-determined dictionaries of words that are associated with the target affect categories are used to generate features for machine learning algorithms. The Linguistic Inquiry and Word Count tool (LIWC) uses a predefined lexicon, counting words in specific psychological categories in order to measure characteristics of a text (Tausczik & Pennebaker, 2010). Taboada et al.'s Semantic Orientation CALculator used manually compiled lexicons for sentiment analysis, with consistent results across a variety of types of product reviews (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011).

Many sentiment analysis techniques are most successful when applied to carefully authored, lengthier content, but often struggle when faced with informal online communication. There have been efforts to adapt some of these techniques to work with such content. Thelwall et al. detected positive, negative, and neutral emotions in MySpace blog posts, which use informal language including nonstandard spellings and grammar. Classifying the strength of both negative and positive sentiment independently on 5 point scales, their algorithm, SentiStrength, performed well relative to other machine learning approaches because of its ability to correct misspellings and its sentiment strength lexicon, in combination with other features (Thelwall et al., 2010).

The Affect Analysis Model uses a database of emoticons, abbreviations, interjections, and other words that were manually associated with nine emotions, to drive a rule-based affect classification system that analyzes affect at the word, phrase, and sentence level (Neviarouskaya, Prendinger, & Ishizuka, 2010). This system was developed specifically for informal online communications, including hundreds of popular abbreviations and emoticons in its database. Over two blog post datasets, the Affect Analysis Model reached 72% and 77% accuracy, and outperformed other systems on news headlines. Although work such as this is promising, spontaneous text communication, such as collaborative online chat, presents other unique problems, with extremely short, hastily written messages, and rapidly shifting topics and affective states.

There are noted challenges in applying lexicon-based approaches, as in LIWC, to naturalistic or jargon-ridden language (Rosé et al., 2008). Some word associations that lexicon-based tools rely on break down in specialized domains. In techniques developed for formal documents, punctuation, nonstandard capitalizations, grammar, and misspellings are often discarded as noise. This makes LIWC and many other existing tools inappropriate in this work. In our dataset, messages often communicate affect through grammar and spelling modification, capitalization,

and punctuation: e.g. “in there???” and “WHAT WHO DID THAT” (*annoyance* and *frustration*). While these issues have been studied in blog posts, very little previous work has focused on affect classification in chat messages.

Besides lexicons, a variety of other approaches have been studied. Liu et al.’s EmpathyBuddy made use of real-world knowledge, obtained from the Open Mind Common Sense knowledgebase, to power an affective email client, which displayed Chernoff face-style feedback alongside email messages. Participants in a user study perceived the email client as more intelligent than a version which displayed random faces (Liu, Lieberman, & Selker, 2003).

Other work has had some success with n-gram features, or short phrases derived from the dataset itself, to inform classification. Rather than a predefined lexicon of words with known associations, the significance of the n-grams is learned from the available data. Aman and Szpakowicz experimented with lexicon-based features from the General Inquirer and WordNet-Affect, in addition to other features, such as emoticons, exclamation points, and question marks, for classifying blog post sentences according to a taxonomy of six basic emotions. The combination of all of these types of features was found to be the most successful configuration, compared to using any one group of features in isolation (Aman & Szpakowicz, 2007).

Mishne supplemented word counts with punctuation, emoticons, and length of blog post. While providing an important exploration of a feature-enhanced analysis approach, the experiment showed a modest 8% improvement over the 50% baseline on average, suggesting that further improvement would be possible with more training data (Mishne, 2005).

Gilbert explored the relationship between the words written in emails and the rank of the email recipient in the workplace hierarchy. Recipients were ranked as higher, lower, or the same as the

sender within a dataset of 2,000 messages from the Enron email corpus. A logistic regression model was used to determine the most predictive unigrams, bigrams, and trigrams from the emails. Support vector machine (SVM) classification based on these phrases achieved 70% accuracy with three-fold cross-validation (Gilbert, 2012).

Learning the terms of interest from the dataset under study can provide both advantages and disadvantages. Mohammad experimented with both lexicons and n-gram features to classify affect in text. Findings showed that the efficacy of word-level lexicons (WordNet-Affect and NRC) was correlated with the size of the lexicon, with the larger lexicon (NRC) showing significant improvements over the use of n-gram features alone for sentence level affect classification. The study also found that the classification performance of n-gram features was domain specific; namely that n-gram features trained on data from one domain were unable to classify affect as well as lexicon features when transferred to a new domain (Mohammad, 2012).

Most work in this area has focused on texts written by an individual, but there has also been work on affect detection in collaborative contexts, e.g. investigating applications of modern text classification techniques for analyzing computer supported cooperative learning environments (Rosé et al., 2008). In this chapter, we address affect classification methods for analyzing chat messages from distributed collaborative work.

## 4.3 Dataset Preparation

Our goal in this project was to automate the process of affect labeling in a specific set of chat logs produced by an extended scientific collaboration. We begin by describing the corpus and the manual labeling process we used to generate truth data for machine learning. Then, we describe the results of experiments on different feature selection and classification configurations.

### 4.3.1 The Supernova Factory Chat Dataset

Our dataset is comprised of chat logs collected from the Nearby Supernova Factory (Aragon, Poon, Monroy-Hernández, & Aragon, 2009), an international astrophysics collaboration of approximately 30 core members; about half of the scientists were located in the U.S. and the other half in France. The scientists were monitoring Type Ia supernovae, a specific type of stellar explosions with a consistent brightness, which allows them to be used to measure the distances to other galaxies and trace the expansion history of the universe. The scientists, distributed across multiple time zones, operated their telescope remotely three nights per week using chat as the primary means of communication; during such operation, numerous technical and scientific decisions involving the operation of the telescope had to be made quickly and collaboratively.

There are a total of 485,045 chat messages in the corpus. The top 32 human participants contributed over 500 messages each, or 300,684 messages total, accounting for the majority of the data. Most of the rest were produced by automated programs ("bots") using chat to relay critical changes to the environment (sunrise/sunset; weather; telescope settings, etc.) (Poon, Thomas, Aragon, & Lee, 2008). Individual chat messages are very short, with the vast majority between 5 and 10 words in length. The chat logs span 1,319 days (nearly four years).

### 4.3.2 Coding for Affect

We prepared training data through a manual coding process. A subset of the data was annotated with any number of non-mutually-exclusive affect codes by between 1 and 5 undergraduate and graduate students in our research group. Below, we describe the mechanics and justification of the manual labeling process.

Because our main goal in this work was to facilitate a rich analysis of the dynamics of distributed work in a specific dataset, we constructed a taxonomy of affect (Scott et al., 2012) through a combination of open, axial, and selective coding grounded in the data (Charmaz, 2006), and affective terms from Plutchik's taxonomy of emotion (Plutchik, 1991, 2001). This approach was our solution to the problem of translating a large body of existing work on affect and emotion into an appropriate and useful analytic tool for our particular dataset. The resulting taxonomy allows us to examine the specific types of affective expression present in our data because it accounts for the distinct ways in which these expressions are molded by the text-based medium. Over the course of several months we conducted iterative, open coding of the chat messages. Coders were allowed to label messages with as many affective terms as they believed applied to each message, and they could add new codes. Messages could also be coded as "no affect."

The interpretation of affect in any recorded communication is inherently subjective. In deciding which affect labels to apply to a given chat message, coders were asked not to attempt to guess what emotion the speaker was feeling, nor the affect that the speaker may have intended to express. Instead, they were asked to focus on the affect that they believed was communicated by the message in the context of the conversation. Coders were encouraged to trust their own instincts as experienced human readers of chat messages from this dataset.

Given the nature of this task, we did not find that any existing qualitative data analysis software was suitable. We needed to allow up to 10 human coders to efficiently work together on thousands of short, sequential lines of chat data, and this data needed to be easily accessible and transformable for machine learning procedures and statistical analysis. Thus, as part of preparing the ground truth data, we developed and used our own web-based coding tool, Text Prizm, pictured in Figure 4.1. The design of Text Prizm is discussed in detail in Chapter 6.

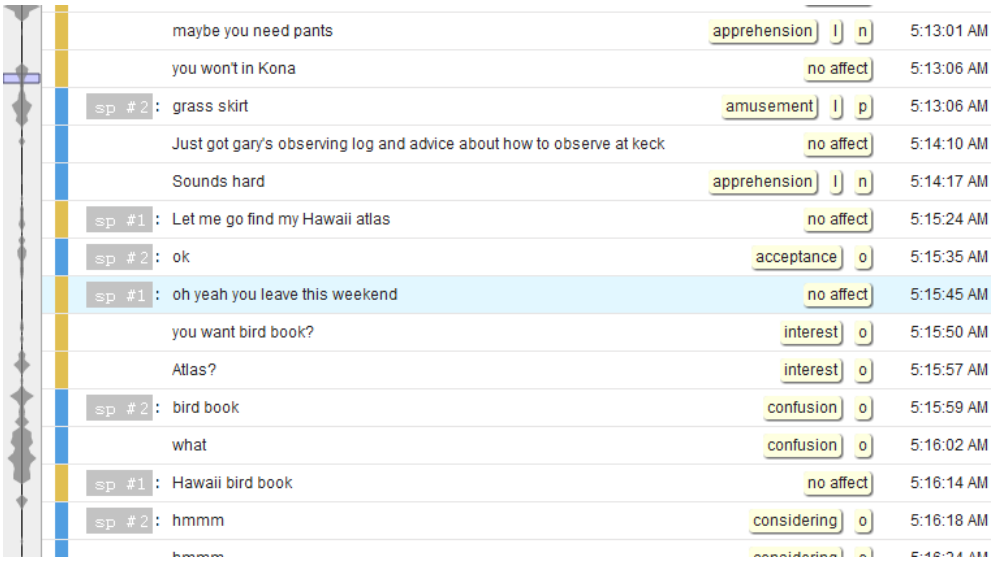


**Figure 4.1. A screenshot of our coding tool, Text Prizm. View seen by an individual coder applying codes. (Codes *o, l, h, n,* and *p* refer to neutral, low, and high intensity, and negative and positive valence, respectively.)**

As the affect taxonomy stabilized, we integrated the categories we had developed with Plutchik's taxonomy of emotion to facilitate comparison to other work and ensure that our taxonomy captured the breadth of possible expressions of emotion. We also added codes for intensity (low, neutral, and high) and valence (positive and negative) to allow for coarse analysis and comparison to prior sentiment analysis research, which typically focuses on positive and negative sentiment. However, we do not address these codes in this chapter. The resulting taxonomy included about 40 different categories. Some categories – such as *interest*, *considering*, and *agreement* – are cognitive aspects that are closely linked with an affective component in the way that they are communicated or expressed by members of the group. This inclusive and flexible aspect of our taxonomy ensures that it captures the broad range of affective expressions that may influence group dynamics.

With the taxonomy solidified, a team of three primary coders and five additional coders, all part of the research group, coded about 5% of the dataset over a period of about 8 weeks. Of 27,344 messages coded, 15,942 (58%) were coded as 'no affect' by at least one person. About 18,000

messages were coded by exactly one rater, the rest by up to 5 raters. Most of the affect codes were used too infrequently for training a machine learning algorithm: Figure 4.2 shows the most commonly applied 13 affect codes, which we focus on throughout this chapter.

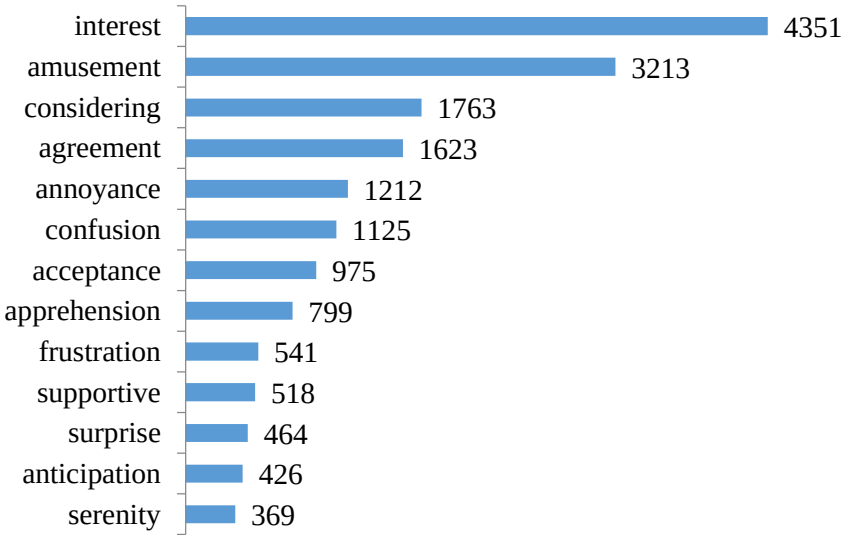


**Figure 4.2. Number of times each of the top 13 affect codes was applied.**

In many cases, multiple affect codes may apply, such as *annoyance* and *frustration* for these three messages sent by the same person: “Did I see a bunch of = vs = = in there??? / WHAT / WHO DID THAT”. These coincident codes sometimes overlap in meaning. At other times, multiple codes with distinct meanings may apply: *anger and confusion* were both applied to a conversation about error-prone software. Examples where this expressiveness is especially useful are included in Figure 4.3. Of the messages coded, 1,599 were coded with more than one affect code by at least one coder, and 129 were coded with more than two affect codes by at least one coder.

| Time | Speaker | Message |
|---|---|---|
| 5:58:41 | Alice | ok, so where was the ******* SN on the image?<br>*#1: interest / anger*<br>*#2: annoyance / confusion#3: interest / frustration* |
| 5:58:55 | Alice | was it the bright blob?<br>*#1: interest / anger*<br>*#2: considering*<br>*#3: interest* |
| 5:59:03 | Ben | 5876 absorption is much wider than the H alpha in v space<br>*#1, #2, #3: no affect* |
| 5:59:18 | Ben | Oh hmmm.<br>*#1, #2, #3: considering* |
| 5:59:28 | Ben | Lemme see what [the] coordinates were...<br>*#1, #2, #3: no affect* |

| Time | Speaker | Message |
|---|---|---|
| 6:13:07 | Charlie | is it “well-developed”?<br>*#1: interest* |
| 6:13:18 | Alice | Should be an interesting experiment.<br>*#1, #2: anticipation*<br>*#3: interest* |
| 6:13:19 | Dana | yes<br>*#1, #3: agreement*<br>*#2: no affect* |
| 6:12:20 | Dana | big!!<br>*#1: excitement / agreement*<br>*#2, #3: excitement* |

**Figure 4.3. Two anonymized examples of conversations from our dataset. Each segment was coded by three members of the research team (#1, #2, #3), below each line.**

### 4.3.3 Reliability

When qualitative coding is used quantitatively, it is common to calculate inter-rater reliability. However, several characteristics of our data complicate the calculation of inter-rater reliability. There are between 1 and 5 raters per message, which can be modeled as missing data, but since the 40 codes of interest are subjective, not mutually exclusive, and may be conceptually overlapping, the criteria for agreement are atypical: non-matching codes may sometimes reflect a degree of agreement instead of absolute disagreement. If we are to estimate the reliability of the

taxonomy as a whole, rather than the reliability for each code independently, then it becomes a problem that raters could apply more than one code per message.

One of the most widely used reliability statistics, Cohen's *kappa* (Cohen, 1960), is not easily applied to this data because of the variable number of coders and the large number of codes. A multitude of variations exist, such as Fleiss' *kappa* (Fleiss, 1971) which can work with more than two coders, but this cannot handle missing data (i.e. variable number of raters). Krippendorff's *alpha* (Hayes & Krippendorff, 2007), one of the most flexible reliability statistics, handles variable numbers of raters, but does not work with non-mutually-exclusive categories.

After studying the available techniques, we decided to modify and extend Cohen's *kappa*. Instead of analyzing the entire taxonomy of affect with its overlapping categories at once, we examined the reliability of each code separately. This also had the advantage of providing useful code-level information. However, we wanted to carefully control the criteria for agreement and disagreement between coders. We also wanted to use all of the data coded by multiple raters, regardless of how many people had coded it. Although Cohen's *kappa* has known weaknesses (Hayes & Krippendorff, 2007), it is also widely understood, so we developed a way to compute *kappa* over our coded data.

In general, *kappa* represents the percent agreement over data points, corrected by the probability of chance agreement. To compute it, we needed to be able to calculate two quantities: the percent of the data points where coders are observed to be in agreement, and an estimate of the probability of coders agreeing by chance. We defined agreement about a particular code on a single chat message in the following way: if more than half of raters said that the code was present, or if all of them said it was absent, then they are in agreement. Otherwise they are in disagreement.

This allows the computation of the percentage of observed agreement, in the manner of Cohen's *kappa*. Computing the probability of coders agreeing by chance directly is more complex because of the variable number of raters and variable number of codes.

We developed an estimate of the probability of chance agreement based on a Monte Carlo method. Similar to the calculations for Cohen's *kappa*, we first calculate the marginal probabilities of choosing each code for each individual coder, and the marginal probabilities of applying specific numbers of codes for each individual coder.

Next, we randomly simulate the rating of a very large number of messages. For each simulated message, we randomly generate the codes that the raters will apply, sampling from the pre-computed marginal distributions. Counting the number of these simulated messages where agreement occurred (according to our definition) allows us to estimate the probability of random agreement. The Monte Carlo simulation continues until all probability estimates are stable to within 0.0001, requiring simulated coding of about 2 million messages.

Finally, we calculate *kappa* in the usual way, given the rate of observed agreement and the probability of chance agreement. Table 4.1 shows the *kappa* values for each of the 13 affect codes that we discuss in this chapter. Our *kappa* values ranged from 0.424 to 0.808, which is generally comparable to reliabilities obtained in previous research coding for affect or emotion (Aman & Szpakowicz, 2007; Neviarouskaya et al., 2010).

| Code | Observed % Agreement | Probability of Chance Agreement | *Kappa* |
|---|---|---|---|
| interest | 0.925 | 0.609 | 0.808 |
| amusement | 0.933 | 0.827 | 0.611 |
| agreement | 0.954 | 0.909 | 0.491 |
| considering | 0.931 | 0.864 | 0.49 |
| confusion | 0.906 | 0.755 | 0.615 |
| acceptance | 0.941 | 0.828 | 0.657 |
| annoyance | 0.929 | 0.693 | 0.77 |
| apprehension | 0.876 | 0.737 | 0.529 |
| supportive | 0.961 | 0.906 | 0.583 |
| surprise | 0.968 | 0.93 | 0.543 |
| anticipation | 0.942 | 0.9 | 0.424 |
| serenity | 0.923 | 0.808 | 0.602 |
| frustration | 0.971 | 0.935 | 0.55 |

**Table 4.1. *Kappa* statistics showing reliability of top 13 affect codes.**

## 4.4 Our Approach

Our purpose in classifying affect is to automatically apply affect labels to our entire chat dataset with reasonable accuracy. With a sufficient amount of chat data coded for affect, we began developing a pipeline of data processing steps and classifier configurations, seeking configurations with the strongest precision and recall for as many different affect categories as possible.

Because most of the affect codes in our taxonomy still had far too few examples to use for machine learning, we focused on only the 13 affect codes that were manually applied over 300 times (Figure 4.2). This included a mix of positive and negative codes, as well as some closely related codes (e.g. *frustration* and *annoyance*).

We have already mentioned several significant challenges to successful classification of chat messages using current text classification algorithms. Facing these problems required numerous choices throughout the classification pipeline, including preprocessing steps, features, and classification algorithms. In order to explore this vast space to find the most successful overall pipeline, we ran a series of experiments designed to test each aspect of the pipeline in isolation. Combining these results yielded the findings we present in this chapter.

We describe the options considered at each step in the pipeline, explaining how we dealt with the challenges specific to chat messages and which choices were supported by experimental results.

## 4.4.1 Experiment Setup

In order to maximize our efficient use of the limited truth data, we used 10-fold cross validation (Witten, Frank, & Hall, 2011) for all experiments. This helps to avoid configurations that work well on training data but do not generalize well. Here we also describe our preparation of truth data and our software architecture.

### 4.4.1.1 Data Preparation

For classification, we transformed the manually-labeled chat messages into training sets, one for each code, with examples labeled “present” or “absent” for each code. In order to do this, we had to decide how to use labels that raters disagreed on, and how to account for the small number of positive examples for each code.

As previously mentioned, raters frequently, and often justifiably, disagree about what affect is present in a given message. For the purposes of creating truth data for classification, we assigned a given affect label in the truth data to any messages where any of the raters applied that affect code. Reliability was low for some affect codes, reflecting low internal consistency within the

raters, but not necessarily low validity (i.e. low inter-rater reliability does not imply that the codes fail to capture affect). Rather, the process by which the taxonomy was developed and applied is the best assurance that the codes we used actually reflect affect expression in the data.

For all of our affect categories, there are far more messages without a given affect code (negative examples) than there are messages where that code applies (positive examples). The imbalance in the labeled data is a common problem in machine learning tasks. In such situations, it can be difficult for a classification algorithm to outperform the baseline (simply guessing the majority class). For this reason, there are many established methods for *balancing* datasets. We experimented with different balancing strategies including up-sampling (randomly duplicating positive examples) and down-sampling (randomly removing negative examples). We found that down-sampling led to more consistent results across different affect codes.

In the experiments reported below, we applied downsampling both to the training sets and to the held-out sets in cross validation. Thus, the performance would be different on real unlabeled data where the percent of positive examples is very low. In order to evaluate classifiers with realistic unlabeled data, the relative importance of minimizing false positives vs. minimizing false negatives must be determined. These decisions depend heavily on the purpose for which labels are needed, and will vary from one project to another. We decided to downsample the held-out sets for this chapter because the results generated are agnostic of the specific project context and are more easily interpreted.

#### 4.4.1.2 Software Architecture

Key to our rapid exploration of different choices for the machine learning pipeline was the use of a relational database for data storage and manipulation. Our web-based chat coding application,

unlike most commercial coding systems, uses a MySQL server for storing both messages and applied codes. Our software for running machine learning experiments connects directly to this database. It is implemented in Java and relies on the popular Weka library for implementations of classification algorithms and feature extraction (Hall et al., 2009).

### 4.4.2 Chat Segmentation

Two significant challenges in classifying chat messages are that the messages are extremely short and highly dependent on context. We developed a segmentation procedure that reduces the number of negligibly small data points and incorporates message context. This section motivates and explains our approach to segmentation.

#### 4.4.2.1 Feature Sparsity

Many approaches to text classification rely on "bag of words" features. The ordered list of words in the raw text is simplified by discarding all the ordering information, leaving only the number of times each word occurred in the document. This approach has been successful for many text analysis problems (Sebastiani, 2002; Witten et al., 2011).

A corpus of text documents generally uses a large number of different words, while each individual document uses relatively few. Thus, the feature space produced by bag of words features has a high dimensionality, but is very sparse since each document has a value of 0 for most of the features. Moreover, the overlap in words between any two documents is likely to be small. Under conditions of sparsity with high dimensionality, machine learning algorithms are prone to overfitting (Domingos, 2012). That is, they might detect and learn patterns in the data that do not generalize beyond the specific group of documents used for training.

With chat messages, this effect is even more pronounced. The average length of messages in our data is 26 characters, or about 5 words. This means that the bag of words created from each data point probably contains only a miniscule fraction of the total vocabulary in the corpus. There are likely to be many spurious patterns created by the random coincidence of these words in messages, making the patterns actually relevant to classification difficult to detect.

#### 4.4.2.2 Context Dependence

Much of the information about individual messages is not present in the messages themselves, but rather in context. As with the context issues reported by researchers in Chapter 3, chat messages are often not understandable without reading many lines before (and sometimes after) in the logs, posing a challenge to message-by-message classification.

In classification problems, the data points being classified are typically assumed to be independent of one another. Yet, chat messages rarely stand on their own. Approaches which do support learning of labels in context of their surroundings, such as hidden Markov models (HMMs) or other graphical models, are available but not commonly used for affect identification in text. We discuss our approaches to accounting for context in the following sections, leaving experiments with graphical models like HMMs for future work.

#### 4.4.2.3 Segmentation Procedure

To deal with small message lengths and to help capture message context, we split the data into segments, combining messages based on their proximity in time. Combination of messages was determined by a simple time threshold: if two messages were separated by less than the threshold, they were grouped together. We evaluated the classification performance achieved with different

thresholds. Because many consecutive messages in our dataset were separated by less than 25 seconds, we tested nine different time thresholds from 5 to 45 seconds at 5 second intervals.

We also developed and evaluated two different formulations of the segmentation procedure. One formulation grouped together messages by different speakers, reflecting an assumption that affective state is distributed among all of the chat participants. The second formulation did not combine messages by different speakers, presuming that affective state is bound to individuals.

Regardless of the time threshold selected, segmentation resulted in a significant reduction in dataset size due to the combination of data points: a conservative threshold of 10 seconds, not separating by participant, halved the number of data points. The effect was less pronounced when we maintained separation between participants: at 10-second-segmentation, the dataset only shrank to two-thirds of its original size. In general, reducing the amount of training data makes machine learning difficult.

However, we also observed that segmentation did have the desired effect: the number of negligibly small messages (40 characters or less, only a few words) in the dataset decreased, having been combined with other messages. We hypothesized that the higher word-count per data point would improve classification results. Additionally, the segmentation procedure makes contextual information available, because the messages in the immediate context of each data point are pulled in and combined.

In our experiments, the effectiveness of segmentation procedures varied from one affect code to another. For some codes, performance differences between different time threshold settings were as great as 10%. For most of the 13 affect codes we tested, the best classification performance was achieved with a time threshold of 30 seconds, keeping messages from different participants in

different segments. We believe that this threshold balances the harmful effects of reducing the dataset size against the benefits of increasing the size of data points.

### 4.4.3 Enriched Feature Space

Face-to-face communication relies on facial expressions and tone to communicate affect. Without these channels, chat participants use other means to communicate affect. We developed a rich set of features to help capture these aspects of chat messages, including pronoun categories, punctuation, emoticons, spelling changes, and the words in the message.

#### 4.4.3.1 Linguistic Challenges

Our dataset, and chat communication generally, is rife with informal language and atypical spelling and punctuation. Many successful techniques for automatically analyzing text, such as LIWC (Tausczik & Pennebaker, 2010), rely on one or more characteristics of standard written language, such as a reasonably correct vocabulary, correct grammar, and predictable punctuation.

For example, in converting a text document into a sequence of words, the text is usually split up at spaces and punctuation characters (although other techniques do exist). Indiscriminate use of this technique on chat communication risks obliterating much of the interesting content. Emoticons and other nonstandard punctuations (*e.g.* "??????!") carry a great deal of meaning in chat, but would typically be removed or distorted during this process.

As noted earlier, lexicon-based approaches to affect detection and sentiment analysis in text (Gill et al., 2008; Mohammad, 2012; Strapparava & Valitutti, 2004; Thelwall et al., 2010) are difficult to adapt to communications with frequent nonstandard spellings and abbreviations. Because of these irregularities and because of their short length, typical chat messages may contain few if any words that are recognized by these tools. In our data, there is also a large amount of jargon and a

mix of multiple languages, making the application of lexicon-based techniques especially challenging.

#### 4.4.3.2 Features

We included duration, length, and rate of messages as features to capture temporal aspects of chat communication. A conversation over a short period of time, with a high rate of messages, could signal urgency or anger, such as a problem with the telescope. Whereas a conversation over a longer period of time and a lower message rate could signal one of the participants explaining a process to another, or a less stressful event.

Grammatical markers, such as personal pronouns, and punctuation, unusual spelling, and emoticons may also communicate affect. For example, the statement "*the telescope is stuck*" can have a markedly different character when expressed using various non-verbal cues embedded in the text (Figure 4.4).

| | |
|---|---|
| "The telescope is stuck! >:(" | Exclamation point and emoticon suggest *frustration*. |
| "The telescope is stuuuuuuuuuck..." | Repetition of the letter "u" suggests *annoyance*. |
| "The telescope is stuck??" | Multiple question marks suggest *confusion*. |

**Figure 4.4. Low-level text features alter the meaning of the phrase.**

In addition to the above features, we included traditional bag-of-words features, based on Weka's *StringToWordVector* filter. Certain words occur more often with specific affect codes. For example, *confusion* is very often linked with phrases containing variations of "confuse," such as "confusing" and "confused." *Amusement* is often paired with phrases containing "haha," as in "You should just live there hahaha" or "Did you have enough coffee this morning? Haha."

#### 4.4.3.3 Evaluation of Feature Sets

We experimented with two different configurations of the bag-of-words features: one using the Porter stemming algorithm[7] (reducing each word to its base form), and one removing stop-words (common words like "and" or "the"). While it reduced the feature space, word stemming had no noticeable effect on classification performance, but the removal of stop-words consistently *decreased* performance by 2-3% for most affect codes we tested. Based on these results, we used a stemming algorithm, but not a stop-list, to generate the bag of words.

In order to determine the value of the other features outside of the bag of words, we measured the performance difference between a dataset prepared with standard bag-of-words features (a large set of about 1.5k words), and similar datasets that were augmented with sets of additional features (such as punctuation, pronouns, or emoticons). Performance improvements of a few percent from each of these new types of features in isolation prompted us to continue developing and improving them, and to combine them into the extensive set of rich features summarized in Figure 4.5. We also applied additional reduction techniques to the bag-of-words features, such as a minimum frequency threshold and lowercasing.

| **Message Information (4 features)** |
|---|
| **duration**: the length of the segment in seconds<br>**length**: the number of characters in the segment<br>**characters per second**: length / duration<br>**rate**: the average rate of messages in the segment |
| **Pronouns (7 features)** |
| # of **1st person singular** pronouns: I, me, my, mine…<br># of **2nd person singular** pronouns: you, your, yours…<br># of **3rd person singular** pronouns: she, he, hers, his…<br># of **1st person plural** pronouns: we, us, ours…<br># of **2nd person plural** pronouns: you all, yourselves |

[7] http://tartarus.org/martin/PorterStemmer/

| |
|---|
| # of **3rd person plural** pronouns: they, them, theirs, their…<br># of **interrogative** pronouns: who, whom, whose |
| **Punctuations (8 features)** |
| # and length of **ellipses**<br># and length of **question marks**<br># and length of **exclamation points**<br># and length of **?!s and !?s** |
| **Special strings (3 features)** |
| # of **negation words**: no, not, cannot, aren't, can't…<br># of **swear words**<br># of **known people names** (list of about 18) |
| **Low-level spelling features (8 features)** |
| # and length of **capital letters**<br># and length of "**hmmm**"-variants: hm, hmm…<br># length of **laughter** phrases: lol, hehe, heehee, haha…<br># and length of **repeated letter sequences** 3 or longer |
| **Emoticons (varies, ~ 8-15 features)** |
| Counts the **number of each emoticon** in the message.<br>Vocabulary of up to about 2200 ( marshall.freeshell.org/smileys.html**)**<br>Emoticons must occur in at least 10 data points to be counted |
| **Bag of Words features (Stemmed & lowercased)**<br>**# of word features varied, ~200-300 words** |

**Figure 4.5. A detailed list of features used.**

### 4.4.4 Classifier Configuration

Aside from the data preparation procedure and the set of features to be used in classification, we considered a variety of options for the classification algorithm itself.

Our taxonomy of affect provides a multitude of categories into which chat messages can be classified. This could be reformulated as a multiclass classification problem, where a single trained classifier would select one from among the 13 affect codes for each message submitted. However, because our categories are not mutually exclusive, we decided to create a separate binary classifier for each of the affect codes tested. We plan to experiment with other configurations in future work.

| Classifier | F-measure | Precision | Recall | Accuracy |
|---|---|---|---|---|
| Naïve Bayes | 0.650 | 0.637 | 0.691 | 0.637 |
| Logistic Reg. | 0.730 | 0.731 | 0.731 | 0.730 |
| **SVM (SMO)** | **0.759** | **0.766** | **0.751** | **0.761** |
| C4.5 (J48) | 0.700 | 0.724 | 0.680 | 0.710 |

**Table 4.2: Performance comparison of classification algorithms. Averaged over 4 runs of cross validation for each of the 13 codes tested, in preliminary experiments.**

| Code | F-measure | Precision | Recall | Accuracy |
|---|---|---|---|---|
| interest | 0.925 | 0.925 | 0.926 | 93% |
| amusement | 0.734 | 0.78 | 0.694 | 75% |
| agreement | 0.779 | 0.813 | 0.748 | 79% |
| considering | 0.761 | 0.774 | 0.749 | 76% |
| confusion | 0.738 | 0.743 | 0.733 | 74% |
| acceptance | 0.773 | 0.805 | 0.743 | 78% |
| annoyance | 0.642 | 0.668 | 0.618 | 66% |
| apprehension | 0.638 | 0.657 | 0.619 | 65% |
| supportive | 0.626 | 0.66 | 0.596 | 64% |
| surprise | 0.71 | 0.789 | 0.645 | 74% |
| anticipation | 0.748 | 0.743 | 0.753 | 75% |
| serenity | 0.663 | 0.74 | 0.601 | 69% |
| frustration | 0.673 | 0.734 | 0.621 | 70% |

**Table 4.3: Classifier performance for each of the top 13 codes. Cross validation was used on data with class frequencies balanced.**

Our early experiments used Weka (Hall et al., 2009) to test a variety of classification algorithms including Naïve Bayes, C4.5 decision trees, support vector machines (SVM), logistic regression, voted perceptron, as well as boosting and bagging. We tested different parameter configurations for each of these algorithms, devoting attention in subsequent experiments to those with promising initial results. In these initial experiments, we found that linear-kernel SVM and logistic regression were quite effective (Table 4.2), which is consistent with prior results (Gilbert, 2012; Joachims,

1998). Our later experimental setups focused on configurations of SVM and logistic regression classifiers, and included Naïve Bayes and decision tree approaches for comparison.

## 4.5 Results and Discussion

In this section we describe the performance of our classification pipeline on each of the 13 most commonly occurring affect codes in our taxonomy (Table 4.3). We also discuss which features were most useful for classifying the affect codes we tested (Table 4.4), issues of methodological alignment between machine learning and qualitative methods, and questions for future work.

### 4.5.1 Classification Performance

The results of evaluating the SVM classifier with 10-fold cross validation for the top 13 affect codes are provided in Table 4.3. Accuracy for most affect codes fell in the 70-80% range. The SVM algorithm is often found to be robust to large feature spaces, which are typical in text classification applications (Witten et al., 2011). Others working on affect classification in text have also found SVMs to be effective (Joachims, 1998; Mishne, 2005; Thelwall et al., 2010).

Additional work is needed on more granular machine learning debugging and development tools that analyze *which* errors classifiers make. It is possible, for example, that one classifier makes qualitatively worse mistakes than another, misclassifying obvious examples, whereas a better classifier only missteps on nuanced cases that are difficult even for humans. This evaluation would require an additional step of human re-evaluation of code appropriateness (Torkildson, 2013).

Furthermore, some apparent classification errors may not actually be errors. Preliminary inspection of a selection of chat messages where the labels produced by the SVM disagreed with our own manual coding found that in many cases a strong case could be made that the classifier's label

actually made sense. As classification methods become increasingly robust and accurate, reaching human accuracy in difficult problems like this one, the errors that human coders make pose a challenge. Although reliability can help to verify the internal consistency of raters, to err is human, and so perhaps classifiers ought to balance measures of confidence against models of human error to produce results more descriptive than precision and recall.

### 4.5.2 Transparency and Interpretability

Of particular interest to researchers are the specific signals that chat participants use to communicate affect. We examined the weight vectors produced by the linear SVM training algorithm to better understand which features were most influential. Note that this type of analysis does not allow quantitative comparison from one trained classifier to another. However, it does reveal the most significant features for each affect code, as modeled by the classifier (Table 4.4).

The top features span many types of features; however, different codes appear to be associated not only with specific features, but also specific feature types. For example, *amusement* is most clearly indicated by various emoticons, while most of the other affect codes do not strongly rely on emoticons. Meanwhile, the *anticipation* code chiefly uses bag-of-words features, words that are often used when discussing the future. In contrast, more immediate, active affect codes, such as *frustration* and *surprise*, are based mostly on punctuation, message rate, and low-level features.

Pennebaker et al. have had success using functional linguistic cues to detect emotional content (Tausczik & Pennebaker, 2010). In everyday speech, there are words we use that carry informational content (the semantics of what we mean to say) and those that make the utterance sensible. The latter, functional elements of text are comparatively more meaning- and context-agnostic. In contrast, EmpathyBuddy relied on real-world knowledge, extracting emotional

content from semantics (Liu et al., 2003). For some of the affect categories that we analyzed, non-semantic cues like capitalization and punctuation appear to be the most important. Semantic cues, including smiley faces and content words that carry meaning, are more useful for other codes.

| **Considering** | **Annoyance** | **Frustration** | **Surprise** |
|---|---|---|---|
| think | *# swearing* | *# swearing* | *# exclamation pts.* |
| *# question marks* | pascal | *# 1st sg. Pronouns* | wow |
| maybe | -- (dash) | *msg. length* | *msg. length* |
| *ellipsis length* | all | *ellipsis length* | *???? length* |
| or | damn | *capital. length* | *!!!! length* |
| *hmm length* | again | *chars/second* | oh |
| *# hmmm* | I | *# negation words* | *ellipsis length* |
| *???? length* | only | it | *# repeated letters* |
| probably | me | *# repeated letters* | *segment duration* |
| x | *msg. length* | *# interrogative prns* | right |
| **Serenity** | **Interest** | **Confusion** | **Apprehension** |
| good | *???? length* | *???? length* | bad |
| : ) | *# question marks* | *# question marks* | something |
| nice | je (fr.) (-) | understand | problem |
| cool | sunrise | confus* | we |
| *!!!! length* | bert | why | seem |
| *msg. length* | est (fr.) (-) | what | too |
| right | where | nothing | *msg. length* |
| too | wonder | wrong | not |
| *# 1st pl. pronouns* | sunset | *msg. length* | *# 3rd sg. Pronouns* |
| do* (-) | interesting | thought | *# swearing* |
| **Amusement** | **Agreement** | **Acceptance** | **Anticipation** |
| ; ) | yes | ok | hope |
| : ) | yeah | okay | if |
| *laughter* | yep | ah | next |
| ; - ) | *msg. length* | *msg. length* | should |
| fun | *segment duration* | *# 1st sg. pronouns* | think |
| *laughter length* | right | oh | will |
| p | yup | yep | try |
| *# people names* | agree | *# question marks* | at |
| sleep | sure | put | like |
| of | okay | *segment duration* | to |

| Supportive |
|---|
| good |
| *???? length (-)* |
| *msg. length* |
| if |
| about |
| the |
| -- (dash) |
| derek |
| he |
| think |

**Table 4.4: Top 10 features for each of 13 classes. (-) indicates that the feature negatively relates to the affect code. * indicates a stemmed word.**

In prior work, it has been common for some categories of affect, e.g. negative sentiment (Thelwall et al., 2010), to be more easily classified than others. The differential weight placed on certain types of features across affect categories suggests that some differences in performance may arise from these researchers' choice of features. In our experiments without emoticons, for example, the classification accuracy for *amusement* decreased significantly. This suggests that future improvements may be obtained by continuing to develop new features, focusing specifically on the worst performing affect categories.

Gill et al. argue that successful social engagement relies on understanding the experience and emotional cues of others, noting the challenge of doing this in a relatively impoverished computer-mediated environment (Gill et al., 2008). Our analysis of influential features for each of the affect categories offers clues as to how different affect states might be expressed and experienced uniquely in text chat. Grammar use, punctuation, the length of responses, and other features, all form a part of this experience much the way facial cues, tone of voice, and body language might augment the emotions of face-to-face communication in different ways. These results suggest that

the experience of affect in text-based chat environments may indeed be much richer and less impoverished than imagined.

In discussing the validity of automating coding, Rosé et al. raise the issue that automated methods base their classifications on the most predictive features, which may not be relevant to the cognitive process of human coding. For this reason, algorithms that have interpretable learned mechanisms are preferred; in some applications, optimizing for quantitative performance metrics is comparatively less important than maintaining a grasp on the internal reasonableness of the classification model itself. We have reported the most predictive features for the SVM classifier, but machine learning algorithms do not necessarily reveal critical information needed by researchers to assess validity (Rosé et al., 2008). Surfacing useful validation information is a challenge for future work.

There are many opportunities to improve the applicability of machine learning and other complex computational analytics approaches for qualitative social science research. However, in the qualitative research community, there is a suspicion of computational tools and methods (Banner & Albarrran, 2009; Coffey, Holbrook, & Atkinson, 1996). Improving the **transparency and interpretability** of machine learning processes through visualizations (Chuang, Manning, & Heer, 2012; Chuang, Ramage, Manning, & Heer, 2012; Talbot, Lee, Kapoor, & Tan, 2009; Torkildson, 2013), as well as through interactive systems (Amershi, Fogarty, Kapoor, & Tan, 2011; Amershi, Fogarty, & Weld, 2012; Brooks et al., 2015; Kulesza, Amershi, Caruana, Fisher, & Charles, 2014; Simard et al., 2014), may help researchers to develop accurate mental models of how the technology works, its strengths and weaknesses, and how best to apply it in qualitative and mixed methods research.

### 4.5.3 Methodological Alignment

Related work by Rosé et al. in the domain of computer-supported collaborative learning has also developed analytic tools that allow people to code data efficiently (Rosé et al., 2008), as we aim to do for analysis of affect in chat messages from distributed collaborations. One key distinction of our approach is the coding scheme itself. Although we did incorporate an existing taxonomy of emotion, we also engaged in a crucial open coding process. This method is especially important for the analysis of communication of a distributed group using an evolving digital medium that influences interactions (Scott et al., 2012). Analyzing the role of affect in distributed collaboration through traces created by technology requires the freedom to refine and expand coding schemes. Most classical supervised machine learning models are trained in batches, but for projects like ours, automation approaches that are adaptable to a dynamic and evolving taxonomy of categories, such as interactive or online machine learning (Blum, 1998; Ware, Frank, Holmes, Hall, & Witten, 2001), may be more appropriate.

However, methodological issues with machine learning and qualitative research can run even deeper, as machine learning techniques may incorporate assumptions that are inconsistent with the researcher's overall methodological approach. Computational and quantitative analyses, including machine learning, depend on certain assumptions about how the world can be modeled and represented. In machine learning, there is typically an assumption that examples can be clearly and unambiguously classified with a true label, and that alternative labels would be erroneous. This assumption is useful because it permits statistical evaluation of machine learning systems.

The classical supervised machine learning problem is formulated in terms of a set of examples or data points, each of which can be described and analyzed on its own, as a unit, independent of

other examples. In the real world, ambiguity and disagreement, contextual dependence, interrelatedness and interaction, and changing circumstances can often complicate matters. While qualitative approaches vary, the interpretive or constructivist lens applied by many qualitative researchers embraces ambiguity, multiple perspectives, and meaning in context (Denzin & Lincoln, 2011). Positivist assumptions of independence, and of a stable, well-defined truth, may be inconsistent with these philosophical commitments, and some qualitative researchers have voiced suspicions about the use of software for qualitative research on these grounds (Banner & Albarrran, 2009; Coffey et al., 1996; Goble, Austin, Larsen, Kreitzer, & Brintnell, 2012). Mixed methods theorists argue that the apparent philosophical incompatibilities between these methods are not as problematic as they seem, and that pragmatic use of multiple methods is both possible and productive (Bergman, 2008; Johnson & Onwuegbuzie, 2004).

In this chapter, we combined a qualitative analysis of affect in text-based chat communication with machine learning analysis. The different philosophical perspectives of these two methods manifested in several ways. Using an interpretive approach, we created training data by manually coding chat messages in a collaborative, open-ended method where multiple emotions and affect labels could be attached to messages. We re-interpreted disagreement between coders as a productive discourse, capturing differing accounts of emotions in the dataset, which we then combined to unambiguous training data using an inclusive heuristic that allowed all votes for each affect label to be counted as true. Chat messages are clearly not independent of one another, but occur in a sporadic, bursty stream full of back-references and conversations. In order to capture some of this interdependence and context in the messages, we grouped messages together into time-proximity-based segments, which were used as the basic examples for machine learning. Beyond the strategies we applied here to integrate machine learning into our mixed methods

project, there are many questions remaining about the role for machine learning in combination with qualitative methods.

### 4.5.4 Learning from Data

Our overall approach here was based on grounded theory, a qualitative social science method that, in principle, privileges data collected from the real world over other sources of knowledge (Charmaz, 2006; Glaser & Strauss, 1967). Several of the interviewees in Chapter 3 also took a grounded approach to analyzing social media data in their research. In grounded theory, the researcher immerses himself or herself in the data, and emerges with a collection of data-originated concepts and relationships (Charmaz, 2006; Scott et al., 2012); care is taken to avoid introducing external theory into the mix, and support for conclusions is taken verbatim from the data. As discussed in Chapter 2, grounded theory, and some other qualitative approaches, privilege the meanings, voices, and perspectives of the specific people being studied. This contrasts with methodologies whose knowledge claims emphasize theoretical constructs from prior research, analytical reasoning or argumentation (e.g. logic and proof), or experimental findings.

Certain types of computational techniques may be more compatible with and supportive of this commitment to real world data than others. For example, a number of recent studies of affect and emotion in online social data have employed off-the-shelf computational analytics packages such as LIWC. LIWC does not *learn* from data; its models have been manually constructed by experts, and validated against datasets from other research sites (Tausczik & Pennebaker, 2010). Tools like LIWC, which come with predefined lexicons or dictionaries of keywords, make the promise of a *generalizable* analysis. For example, LIWC's "positive affect" score, aims to measure affect in a general, cross-domain sense, because its dictionaries are based on decontextualized common

English. These techniques are not sensitive to local aspects of the dataset being analyzed, such as unusual vocabulary, word usages and meanings, spelling, or punctuation patterns (Rosé et al., 2008). In the philosophy of grounded theory, where localized data and knowledge is primary, this becomes a problematic limitation.

This suggests that **corpus-trained machine learning or hybrid learning** approaches may be more suitable for grounded theory-based analysis of online social data. Some researchers have developed hybrid computational approaches that combine predefined lexicon systems (e.g. LIWC) with trained systems such as logistic regression classifiers (Quercia, Ellis, Capra, & Crowcroft, 2012), or latent-semantic analysis (LSA) (Gill et al., 2008), a corpus-based topic modeling technique. In this chapter, we created a machine learning solution that uses features and models derived entirely from the data being analyzed, and our models were evaluated using test data from our corpus. This approach is more likely to be considered sound and sensible within a grounded theory framework. While few of the interviewees in Chapter 3 had tried any form of supervised classification on their social media data, they did mention a variety of unsupervised clustering techniques that they used in exploratory analysis to find structures in the data.

## 4.6 Conclusion

In this chapter, I have discussed an application of machine learning to scale fine-grained, subjective human analysis up to a large chat log. We interpreted affect as a dynamic phenomenon by segmenting the chat dataset on a temporal, per-participant basis, and augmented a standard bag-of-words feature set with analogues of non-verbal cues, such as grammatical markers including unusual spelling and emoticons, and meta-information about the chat messages such as duration,

length, and rate. These decisions led to better classification results, though other avenues of exploration may offer additional improvements.

We were able to classify text for 13 affect codes with F-measures of 70-90%. We also produced a set of predictive features for each of the 13 types of affect considered, which may be applicable in other domains. We open sourced our machine learning software, *ALOE*, to facilitate validation, comparison to other techniques, and promote future research on affect within the community.

Because of the scale and complexity of online communication datasets, a range of computational techniques (e.g. machine learning, clustering, topic modeling, computational linguistics) have been created and deployed by social scientists to address various research questions. However, some techniques may be more readily applied than others. Any given computational technique, as an applied component of a research method, may be aligned or misaligned with the overall methodological orientation of the researchers, and researchers applying automated or partially automated analysis to qualitative research may struggle to reconcile these misalignments. Based on our experience building ALOE for studying affect in chat logs, we have reflected on several potential misalignments between computational techniques with qualitative research practices. This is meant to highlight potential pitfalls and threats to validity for researchers who would apply techniques such as machine learning in their mixed methods projects. Further research is needed to explore additional opportunities for building bridges between qualitative analysis and large-scale computational and quantitative methods.

# Chapter 5: Visual Exploration of Social Media Data

Researchers in the social sciences use datasets collected from social media sites to address a wide variety of research questions, from public health (Paul & Dredze, 2011) and happiness (Dodds, Harris, Kloumann, Bliss, & Danforth, 2011), to political activism (Agarwal, Bennett, Johnson, & Walker, 2014) and disaster response (Starbird et al., 2015). Many more examples are reviewed in Chapter 2. Social media datasets can be quite large, and have complex network structures that shift and change over time. The main content of social media datasets is text-based messaging, which is difficult to analyze at a large scale. As discussed in Chapter 3, these characteristics make social media datasets challenging to work with; even getting a good overview of a social dataset can be very difficult. Researchers often explore a multitude of tools and techniques to understand their data from many perspectives.

Interactive data visualization can enable researchers to reach new insights, and several general-purpose visual analysis tools such as Tableau, Microsoft Excel, and Google Spreadsheets, are commonly used in practice. However, these tools require considerable data manipulation and preprocessing because they do not easily support social media data. General-purpose visual analysis tools often do not provide integrated support for multiple aspects of social media data, fluid zooming across levels of time and geography, and presentation of social media context. The visual analytics and human-computer interaction literature includes several examples of visualization systems for social media data, such as *Vox Civitas* (Diakopoulos, Naaman, & Kivran-Swaine, 2010) and twitInfo (Marcus et al., 2011). However, here has been little work on exploratory social media visual analytics focused on addressing the needs of social scientists. Because social media research projects in the social sciences are often interdisciplinary,

collaborative tools may be particularly useful, but previous research has not investigated the potential of *collaborative* visual analysis tools for social media data.

In this chapter, we explore the question of how visual analytics tools should be designed to help social science researchers more easily explore large social media datasets. We discuss our research group's experience designing and evaluating Agave, a collaborative visual analysis system for exploring events and sentiment over time in Twitter datasets, with a special focus on sentiment. In Agave, timeline visualizations of trends and spikes in sentiment help teams of users find relevant events, which can be examined in greater detail through filtered lists of tweets. Annotation and discussion features allow users to collaborate as they explore the data.

We recruited a group of researchers to evaluate Agave by exploring a dataset of almost 8 million tweets from the 2013 Super Bowl, a major televised sports event in the United States. We contribute the findings of our qualitative study, discussing the usefulness of collaboration for exploratory analysis of difficult social media datasets, and implications for the design of sentiment visualizations. Agave and its source code are publicly available[8] to encourage further development and research on collaborative social media analysis tools and sentiment visualization.

## 5.1 Acknowledgments

The Agave visual analysis tool discussed in this chapter was designed, implemented, and evaluated in collaboration with John J. Robinson, Megan K. Torkildson, Ray Hong, and Cecilia R. Aragon. A shorter version of this chapter was previously published as "Collaborative Visual Analysis of Sentiment in Twitter Events" at the 2014 conference on Cooperative Design, Visualization, and

[8] http://depts.washington.edu/sccl/tools

Engineering (Brooks, Robinson, Torkildson, Hong, & Aragon, 2014), and then expanded and edited by the primary author.

## 5.2 Background and Related Work

Below, we briefly review examples of Twitter research focused on emotion and sentiment to provide context for how Agave might be used. Following this, we discuss related work on visual analysis of Twitter data and collaborative visual analysis.

### 5.2.1 Emotion in Twitter

While researchers typically analyze multiple aspects of social media data, the research community has demonstrated a strong interest in emotion, affect, and sentiment as it is captured in social media and online communication. Tweets are often explicitly emotional, or carry emotional connotations. Dodds et al. demonstrate how Twitter can be used to calculate a metric for social happiness, and analyze temporal fluctuations in happiness on Twitter over days, months, and years (Dodds et al., 2011). In a similar vein, Quercia et al. calculated a gross community happiness metric based on tweets originating from different census communities in the UK, finding that their metric correlated with socio-economic status at the community level (Quercia, Ellis, Capra, & Crowcroft, 2012). Mood extracted from Twitter has also been associated with daily changes in the stock market (Bollen, Mao, & Zeng, 2011).

At a personal scale, a study of individual tweeting behavior has associated sharing emotion in tweets with having larger, sparser follower networks (Kivran-Swaine & Naaman, 2011). De Choudhury et al. have used mood patterns in the social media activities of individuals to understand behavior changes related to childbirth (De Choudhury, Counts, & Horvitz, 2013), and to recognize signs of depression (De Choudhury, Gamon, Counts, & Horvitz, 2013).

More examples of social science research with social media data were reviewed in Chapter 2, and the need for more exploratory data analysis tools for social media data was discussed extensively by participants in Chapter 3. Our goal in this paper is to explore how collaborative visualization tools can support data exploration in social media research projects.

### 5.2.2 Visual Social Media Analytics

Visualization and visual analytics are promising tools for tackling complex, dynamic social media datasets, and collaborative analysis can enable research teams to reach greater insight in interdisciplinary projects. Several visual analytics systems for social media data have been created in the research community. The "Visual Backchannel" system developed by Dork et al. presents a stream graph of Twitter topics over time, as well as a display of relevant Twitter usernames, tweets, and photos (Dork, Gruen, Williamson, & Carpendale, 2010). The *Vox Civitas* (Diakopoulos et al., 2010) and twitInfo (Marcus et al., 2011) systems use temporal visualizations and sentiment analysis to support journalists exploring tweets about specific events. Hubmann-Haidvogel et al. use dynamic topographic maps to how social media coverage of climate change changes over time (Hubmann-Haidvogel, Brasoveanu, Scharl, Sabou, & Gindl, 2012). For emergency responders, Mazumdar et al. focus on visualizing contextual information in social media data (Mazumdar, Ciravegna, Gentile, & Lanfranchi, 2012). However, these projects did not address collaborative visual analysis, or focus attention on the needs of social scientists.

Over the past 10 years, there has also been interest in collaborative visual analytics in general. The NameVoyager system was an early example of large-scale public data analysis (Wattenberg & Kriss, 2006). Heer et al. followed up on this idea with additional exploration of collaborative features such as graphical annotation, view sharing, and threaded comments (Heer, Viegas, &

Wattenberg, 2007). Focusing on analyst teams, Heer & Agrawala presented a summary of design considerations for collaborative visual analytics systems (Heer & Agrawala, 2008). However, there is little related work at the intersection of visualization, collaborative data analysis, and social media studies, especially focusing on sentiment.

## 5.3 Design of Agave

In this section, we discuss our design process and rationale for the Agave visual analytics tool.

### 5.3.1 Design Process

Agave was created in 2012 and 2013 by myself and a group of researchers who had previously collaborated on the projects discussed in Chapters 4 and 6. Our prior experience working together to analyze affect in text-based chat communication led us to form a new team focused on finding ways to visualize dynamics and transmission of emotional content in the larger information space of Twitter. The mixed background of our group included visualization design, programming, and data science, as well as psychology, information science, and human computer interaction. Early in the project, we collected several large datasets, such as the Super Bowl dataset discussed later in this chapter, and began brainstorming ideas for visualizations.

We quickly encountered problems similar to those discussed by interviewees in Chapter 3. The Twitter datasets were large, and were far more diverse and volatile than we expected. It took considerable effort to understand what was going on in the data even at a high level, let alone to understand (and visualize) a specific, subtle social psychological phenomenon. Before we could pursue our original interest in visualizing the sharing of emotion in messages, we found ourselves creating many simpler visualizations and statistics summarizing the dataset, to understand its peaks and valleys, turning points, and topic streams.

Eventually, we refocused our attention on this new challenge: exploring social media datasets collected during major events (emphasizing the importance of temporal perspectives on the data). Over a period of several months, we conducted extensive brainstorming, sketching, and prototyping. While we started out working primarily on a standalone visualization, we realized that a larger collection of coordinated views, both visual and textual, was needed to understand our datasets. We gradually progressed towards a more full-featured visual analytics tool. As a continuation of our original interest in emotion in Twitter, and because sentiment is a common lens for analyzing social media data (Go, Bhayani, & Huang, 2009; Hao et al., 2011; Zafarani, Cole, & Liu, 2010), we maintained a special focus on automatically-extracted sentiment data.

We also realized that designing the tool to support collaboration could greatly increase its usefulness. Our design process for Agave was motivated by our own experiences trying to explore data as a collaborative research team; research on Twitter is often conducted by such groups because multiple skillsets and disciplinary perspectives are needed. We designed Agave to support communication and collaboration between group members.

The target users for Agave are multidisciplinary teams of researchers or analysts who have collected tweets about a particular event, or topic, over time. For these users, time is a key dimension for understanding the event, and the role played by sentiment, affect, and emotion is of special interest. As we have discussed above, sentiment and emotion are becoming important tools for analyzing social media data, and we have developed several ways of surfacing this information in Agave. Given a newly-gathered set of tweets, the researcher's first need is to explore it to see what it contains, and this is the purpose of Agave. The interface is illustrated in Figure 5.1.

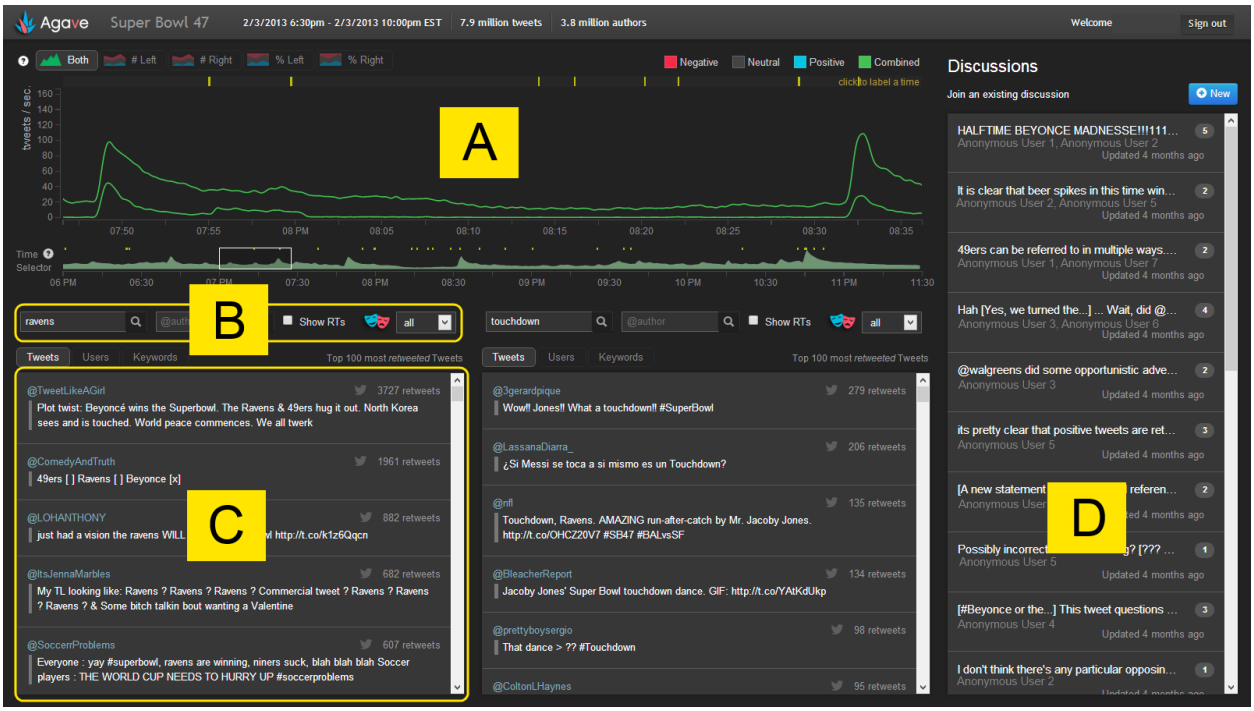


**Figure 5.1: Agave interface, with four main features highlighted. (A) Timeline visualizations displaying different representations of the data, (B) data filters to refine searches, (C) data details showing lists of tweets, users, and keywords, and (D) threaded discussions to communicate with other users.**

### 5.3.2 Timeline Visualizations

Agave helps the user focus on temporal trends through a prominent timeline view. We employ the *focus+context* paradigm (Bjork & Redström, 2000), dividing our timeline visualization into a main view and a smaller "overview" timeline just below it. This separation enables the user to view the rate of tweets at any temporal scale. This allows more data to be displayed on the screen and helps the user maintain context by showing how short-term changes fit into more gradual trends. With the overview, the user can brush a time interval by clicking and dragging to highlight an interesting section. The selected region can also be clicked and dragged, to pan through the dataset.

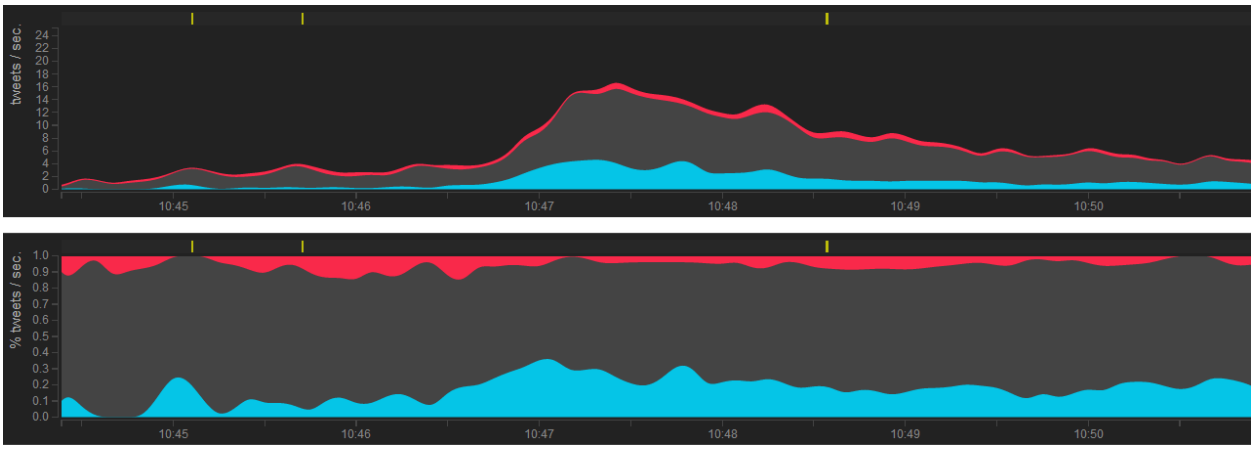


**Figure 5.2. Sentiment streamgraphs for a keyword search. Negative is red, neutral is gray, and positive is blue. *Top*: overall frequency of tweets, divided by sentiment type. *Bottom*: sentiment as percent of overall volume.**

To help users find interesting events, we have included multiple display modes for the timeline. In the default mode (Figure 5.1, A), the timeline shows the *rate of tweets* over time as a line graph. To analyze changes in *sentiment* over time, the other modes show a stacked *streamgraph*-style visualization, similar to ThemeRiver (Havre, Hetzler, Whitney, & Nowell, 2002), with positive, negative, and neutral layers. In one mode, the layers represent tweet counts (Figure 5.2, top), making it easy to see sentiment categories in relation to the overall tweet volume. In the second mode, the layers represent percents (Figure 5.2, bottom), which highlights relative changes in the distribution of tweets. Both modes are useful, and have different strengths and weaknesses.

### 5.3.3 Tweet Details and Filters

When an interval of time is selected using the timeline, the two tabbed panels below the timeline are updated to display the top 100 most-retweeted tweets and users for the selected time interval (Figure 5.1, C). These detail panels provide a snapshot of the twitter activity within the selected time range, answering the question “what was happening during the time interval I selected.”

In addition to tweets and users, the detail panels can also show "burst keywords," which are keywords that rapidly increased in usage during the selected time range. Burst keywords are derived using the technique developed by Guzman and Poblete, involving binning the entire dataset into 5 minute windows and then tallying the number of occurrences of each word in each bin, excluding common stop words (Guzman & Poblete, 2013). We display the 100 keywords with the largest percent increases from all 5 minute intervals that intersect with the selected time range, sorted by percentage increase. However, because the percent increase can sometimes result in a counter-intuitively high ranking for keywords with few occurrences, we also display the number of times each keyword was used and the absolute increase in uses.

To help users dig deeper, the detail panels provide several filter options: search by text and account name, whether to ignore retweets (40-60% of tweets collected are typically retweets), and filter by sentiment. Setting filters will further restrict the example tweets shown in the panels below, and will also restrict the set of tweets plotted in the timeline visualizations above. Comparison of multiple filtered sets of tweets was also a crucial task for our group, and enables the construction and exploration of more powerful and refined questions about the dataset. We included two parallel sets of filters for compare and contrast tasks. Figure 5.1 shows a configuration with two different search queries being compared in the detail panels and on the visualization.

### 5.3.4 Collaborative Discussion and Annotation

To help teams collaborate in the exploration of Twitter data with Agave, we developed two collaborative data analysis features: annotation and discussion threads.

Agave allows users to attach annotations to the tweet timeline through the "annotation bar," a region just above the timeline which can be clicked to add an annotation and comment. Annotation

of visualizations has many benefits, including communicating details between users and providing context in threaded discussions (Heer et al., 2007). Because complex events in Twitter often consist of many smaller events, it is useful to mark these events in Agave once they have been discovered, so that other users can also benefit from these discoveries. For example, the Super Bowl dataset we use here includes many smaller events, such as the start and end of each quarter, touchdowns (scoring points), the halftime show, and commercial breaks. Annotations can be added to mark these significant points in the dataset.

To help teams using Agave write about the data, share their findings with others and cite specific evidence related to their findings, we included a threaded discussions feature. Previous research has found that discussions forums can be useful for collaborative data analysis (Wattenberg, 2005), especially when they are linked to and from the data visualizations (Heer et al., 2007). Willet et al. demonstrated that threaded discussions side-by-side with interactive visualizations made evidence-gathering and consensus-building effective (Willett, Heer, Hellerstein, & Agrawala, 2011). Through a panel on the right side of the interface, users can view and post to threaded discussions (Figure 5.1, D); the most recently updated threads are shown at the top. When users click on a thread, they can see all the existing messages within that thread and post new messages.

To make the threaded discussions more powerful for talking about the data, users may attach explicit references to specific tweets and annotations in their discussion posts. To provide common ground for discussions, it is also useful to preserve and share context in discussion posts. When a user creates a post, Agave saves the current state of the timeline selection, view type, and filters; Agave attaches this state information to the post. Other users can then click a simple link to reconstruct the poster's view of the data and understand the context for the comment. This has been found to accelerate communication and reduce ambiguity (Willett et al., 2011).

## 5.4 Evaluation

We evaluated Agave to investigate how collaborative features and sentiment visualizations could support exploratory analysis. Below, we discuss the Twitter data we used in the study, and the study procedure.

### 5.4.1 Twitter Data Collection

We collected a set of almost 8 million tweets during Super Bowl XLVII, the 2013 annual championship football game of the U.S. National Football League. This event was selected based on an expectation of high Twitter volume with emotional content. Data was collected from the Twitter Streaming API, using a tracking list of 142 terms including team and player names, coaches, and entertainers, from Friday, February 1st at 22:30 EST until 20:30 EST on Tuesday, February 5th.

For sentiment analysis, we used the Sentiment140 API (Go et al., 2009), which categorizes each individual tweet as positive, negative, or neutral. For validation, two researchers in our group manually labeled 500 randomly selected tweets, achieving a Cohen's kappa of 0.57. Compared to these labels, Sentiment140 achieved an overall accuracy of 71%, and a Cohen's kappa of 0.26.

### 5.4.2 Procedure

Because of the open-ended nature of exploratory data analysis, our evaluation used a qualitative approach. Participants explored real Twitter data in Agave and posted comments and annotations. We focused on the following research questions:

- How do the collaborative features support analysis?
- How useful are the sentiment visualizations and filters?

- What types of findings does Agave help generate?
- What problems do users encounter with Agave?

We recruited 7 participants experienced with Twitter-based research (one was also later interviewed for the ethnographic study in Chapter 3). Participants' prior grounding and interest in social media data analysis supports the validity of their interpretations of the dataset and their usage of Agave. After a 5-10 minute tutorial on Agave, participants explored the tool freely for 20-30 minutes, some in our lab, and others remotely by video conference. We used a think-aloud protocol and observer notes to monitor these open-ended sessions, similar to the lab study used in (Heer et al., 2007). We then allowed participants 3-4 days to revisit Agave to continue exploring the dataset on their own. After this out-of-lab session, participants completed a questionnaire about discoveries, problems encountered, and attitudes about Agave. Log data were also collected by the system to determine how often each feature was used. Finally, post-study interviews focused on how participants used the visualizations and collaborative features to explore the data.

## 5.5 Findings and Discussion

Below, we discuss how participants used the timeline and example tweet displays to make sense of events in the Twitter data, the importance of indirect collaboration for exploring the data, and the interpretation of sentiment visualizations.

### 5.5.1 Foraging for Information

Participants used the complementary affordances of the timeline and the tweet list in an iterative, non-linear, information foraging or sensemaking loop (Klein, Moon, & Hoffman, 2006; Pirolli & Card, 2005; Russell, Stefik, Pirolli, & Card, 1993). Participants found events of possible significance through the timeline, and read tweets during the event to find out what the event was

about. Having noticed topics of interest in the tweets, they used the keyword filters to focus on changes over time, which often revealed still more events. We illustrate these processes with examples, below.

The timeline visualizations were an important point of entry into the dataset, since the display of tweet rate over time, optionally divided into color-coded sentiment layers, made it clear where there were sudden changes in tweet volume or sentiment. Participants interpreted these changes as points worthy of further investigation. Below, Peter reflects on his use of the timeline:

> I would say I find peaks or valleys, mostly peaks, then scrunch the time filter around that, then look at the [tweet list] to see if I could find some theme.
>
> – *Peter*

In this example, narrowing the time filter to surround the possible event allows the user to read tweets concerning only that interval of time, providing a hypothesis for what event in the Super Bowl caused the peak or valley. The Super Bowl XLVII was unusual in that it was interrupted for 34 minutes by a power outage. Based on interacting with the timeline, Allison was able to discover that a change in tweet rate during the second half of the game was probably caused by the power returning to the stadium:

> I see some really distinct spikes, and I'm wondering what this is about. […] I want what the tweets are at this time, so I'm going to try making the window really small, and I'm reading the tweets. […] Apparently the lights just came back on, because of this tweet... [pointing to tweet]
>
> – *Allison*

Both of these examples show how the timeline led participants to narrow their focus onto the tweet lists. However, users also went in the other direction, where examining the tweet list prompted a closer inspection of the broader timeline. Having zoomed in on an event, Allison discovers tweets about the Puppy Bowl (a simultaneously broadcast puppy-based alternative to the Super Bowl):

> I see a tweet about the Puppy Bowl [...] so I'm seeing how much you have about that in your dataset. […] I see fewer tweets about it, however, and I see lots of these Puppy Bowl ones reference the Super Bowl. [searches for "puppy bowl"] I see very different spikes for the Puppy Bowl than for the Super Bowl.
>
> – *Allison*

As in this example, when an interesting topic was discovered through the tweet list, participants often searched for tweets referencing that topic, and would then see how the rate of those tweets varied over time. For a few participants, the list of users and burst keywords also revealed potential topics to search for, but the tweet list was used much more often.

The above example also highlights the utility of the paired left and right filters. Allison was able to see the timeline for "puppy bowl" superimposed directly over the timeline for "Super Bowl". Several participants used this feature to compare tweets about the two teams that were competing in the Super Bowl:

> I compared the "49ers" signal against the "Ravens", to see the behavior of the positive, negative, and neutral tweets, over time... the kick-off, half time, and the

> end of the game. Another filter that I used was trying to see the behavior of the positive and negative signals when the blackout occurred.
>
> *– George*

Another participant used the left set of filters to search for touchdowns and the right set was set to show swearwords. While being able to compare filters was certainly useful, some participants found it confusing to keep track of which timeline belonged to which set of filters, suggesting that the design could be improved.

In general, participants switched frequently between the tool's controls, exploring the dataset in an opportunistic hunt for interesting information. Each participant followed a different trajectory through the data and discovered different events and topics that they were interested in. The patterns of interacting with the visualization, searching, and reading raw data, which we observed in Agave, resemble the exploratory analysis practices described by interviewees in Chapter 3.

### 5.5.2 Indirect Collaboration

The design of our evaluation prevented participants from engaging in deeply collaborative analysis. Because the participants did not use the tool at the same time, and did not know each other, there was little reason for them to expect direct responses to their discussion posts. However, while participants in our study only occasionally addressed other users directly, we still observed indirect interactions between people that nonetheless constitute collaborative analysis.

To get an overview of how participants used the collaborative features of Agave (e.g. annotations, discussions, link references to the data), we analyzed the log data collected during the study, summarized in Table 5.1. There were 16 discussion threads created, with a median of 2 replies in

each thread. Topics discussed included opportunistic advertising, retweet patterns, and the halftime show. Six of the seven participants used annotations to mark specific points within the timeline. Some of these annotations included sentiment spikes, fumbles, touchdowns, and advertising-triggered conversation spikes. Participants' findings were typically about specific events represented in the tweets, such as touchdowns.

| Participant | Annotations | New Threads | Posts | Links |
|---|---|---|---|---|
| Krista | 2 | 5 | 10 | 14 |
| Jack | 5 | 3 | 6 | 0 |
| Lauren | 1 | 2 | 4 | 1 |
| Allison | 3 | 2 | 5 | 4 |
| Hannah | 2 | 1 | 3 | 0 |
| Peter | 3 | 2 | 2 | 2 |
| George | 0 | 1 | 1 | 0 |
| **Total** | **16** | **16** | **31** | **21** |

**Table 5.1: Collaborative content created by each participant. The initial post used to create a discussion thread is included in the post count.**

Threads in the study were short (with fewer than 5 posts), and some of these discussions had only one participant. For example, one participant appropriated discussion posts as a way of bookmarking interesting tweets. Of course, these posts were still public, so it was possible for any other user to view and interact with them. As a result, even posts created without any expectation that others would respond were often read by other participants:

> I am looking at a discussion post [in a thread about misclassified sentiment]: "#IRegretNothing is a positive hashtag..." I wonder how many people used that. [Searches for the hashtag] Now I'm looking at RTs for #IRegretNothing.
>
> *– Peter*

This indirect interaction through the discussion threads was useful for suggesting and extending lines of analysis. Participants read comments made by other users and sometimes restored other users' views so that they could easily focus on a potentially interesting section of the data, often adapting and extending the filters created by others.

Even more than discussion posts, annotations created by previous users provided "jumping off points" for exploration of the data. Allowed to explore the tool freely, a few participants were initially unsure how to get started with the Twitter data, and began by opening the discussions or examining the annotations that had been left by other participants. Many participants seemed to prefer the annotations:

> The discussions I found harder to look at and understand than just the yellow tags [annotations], which were fairly straight forward, and I looked at a bunch of those.
>
> – *Peter*

Participants used existing annotations, which were shown in context as part of the timeline, as a way of choosing periods of time to focus on, i.e. social navigation (Dourish & Chalmers, 1994). Just as participants noticed and zoomed in on sharp changes in the timeline, several participants focused on times surrounding annotations in order to understand what the annotations were referring to. The following excerpt from our observer notes shows how Peter used the annotations created by other users to understand how the 49ers fans reacted to their team's defeat:

> Peter focuses the timeline on an interval surrounding the end of the game, where another user's annotation has drawn his attention. He then uses the sentiment filter

> to limit the results to negative sentiment. Browsing through the details pane, he notices multiple tweets about the 49ers losing.

In this case, another participant's annotation at the end of the football game highlighted and explained a shift in sentiment visible in the timeline display. Sometimes, instead of zooming in on annotations, participants sought out relatively unexplored areas of the dataset to focus on, an example of "anti-social navigation" (Heer & Agrawala, 2008; Wattenberg & Kriss, 2006):

> I started by kind of looking in the trough of the blackout period because that hadn't had a lot of annotations, so I thought it would be interesting to look there. In that dip during the blackout, I saw that [a company] was tweeting and got retweeted a bunch of times.
>
> – *Hannah*

Participants expressed appreciation for the annotations feature in particular; as a low-cost way of indirectly collaborating with other users, annotations were valued for helping people start to explore the data. These observations support Heer et al. who argued that "doubly-linked" discussions (linked to and from the relevant data or view) enable users not only to tell what data others are talking about, but also to find existing discussions about data they are viewing (Heer et al., 2007). Our participants noticed annotations above the timeline, which they followed by exploring the time interval around the markings, but Agave did not fully close the "doubly-linked" loop by connecting these annotations and the Twitter data to the discussion posts where they were referenced. Still, our findings suggest that this is a promising idea for future work:

> My goal was very data driven, starting with the data, and there's no real way to end up in the discussion if you start with the data. [...] I think being able to maintain

> links to the data (narrative centric) vs. marking the data itself (data centric) is really great.
>
> – *Krista*

As Krista pointed out, it was difficult to transition from a view of the data to a relevant discussion. Several other participants in the study also pointed out ways that the collaborative side of Agave could be better integrated with the exploratory, analytical views. The complexity of designing and engineering a collaborative visual analytics tool which effectively integrates "narrative-centric" and "data-centric" views remains a barrier to widespread development and adoption of these powerful techniques. Heer & Agrawala have documented the design challenges in some detail (Heer & Agrawala, 2008). Future work on this problem might include additional codification of design patterns for collaborative data analysis tools, and the creation of libraries or reference implementations to demonstrate successful designs. Towards that end, we have released Agave and its source code for exploration and extension.

### 5.5.3 Exploration of Sentiment

Emotion, affect, and sentiment are important facets of social datasets (Diakopoulos et al., 2010; Dork et al., 2010; Marcus et al., 2011). However, the implications of incorporating sentiment into visualizations have not been thoroughly considered. Agave presents information about sentiment in the timeline visualizations and beside tweets in the details panels (Figure 5.1, C).

Although the visualization of total tweets over time was useful for identifying events of interest, some events were more easily identified using the visualizations of positive, neutral, and negative sentiment (Figure 5.2). Peter reflected on his use of the sentiment timeline: "I think I was most interested in visually seeing where the peaks of positive and negative were and what was clustered

around those areas." The sentiment timelines successfully facilitated insight about particularly emotional topics or events such as touchdowns or advertisements, but the sentiment indicators that we attached to individual tweets in the tweet list provoked doubt and suspicion about the validity of the sentiment data:

> I saw a few more tweets in there that were in the positive or negative column which were questionably positive or negative and which I felt were more neutral.
>
> – *Allison*

While sentiment classifications appeared to be merely "wrong," they are actually fundamentally ambiguous and subjective. Boehner et al. have criticized affective computing research which uses an "informational model of emotion," which reduces emotionality to an internal, individual state which can be therefore be measured, modeled, and represented as information. The interactional approach which they propose instead conceives of emotion as cultural, dynamically experienced, and constructed in action, challenging any attempt to model it as information (Boehner, Depaula, Dourish, & Sengers, 2007). The classification of sentiment expressed in tweets as positive, neutral, or negative clearly draws on an informational approach. The cost of this simplification is that, whatever the method of sentiment analysis, the results will always be subject to debate.

In our study, the small sentiment indicators attached to individual tweets in the tweet list made participants question the sentiment data. While this appears problematic at first, sometimes it may be desirable for users to question the uncertainty and subjectivity of their data. There is a risk, with computational analytics such as sentiment analysis, for users to accept the output without question and therefore potentially to form unjustified conclusions based on the visualizations. Rather, they may benefit from accurately understanding the limitations of the data they are working with.

Additional research is needed to understand how to balance trust and validation in visual analytics systems that rely on questionable and ambiguous data.

## 5.6 Conclusions and Future Work

In this chapter, I have presented the design and evaluation of Agave, a collaborative visual analytics tool supporting exploratory analysis of events in large Twitter datasets, with a focus on sentiment. We conducted a qualitative evaluation to find out how researchers used Agave to develop insights about an 8 million tweet dataset. Participants found the timeline visualizations, particularly displays of sentiment changes over time, useful for finding and understanding interesting events. Annotation and discussion helped participants share findings, but also enabled indirect collaboration that stimulated broader and deeper exploration of the data.

The interviews with social scientists in Chapter 3 revealed several key implications for software to support analyzing social media data. According to my interviewees, exploration of social media data is an important and time-consuming part of their research process, and involves examining many aspects of the data, developing a sense of context around the data, and switching between different levels and units of analysis. Agave makes use of several social media fields, including the text (both as sentiment and as raw data), timestamps (for generating timelines), account usernames, and retweet counts (for ranking tweets to display). Integrating these key aspects of the dataset reduces the burden on the user to generate their own separate analyses and queries, providing an efficient user experience for social scientists. While Agave does not incorporate external contextual data, its timeline visualizations show how specific keywords evolve over time, and its search results let users see those keywords in context. In this sense, Agave makes it easy for users to examine specific words in context of the entire dataset. Finally, Agave's timeline

visualization provides smooth zooming and automatic re-aggregation of the data at an appropriate granularity, e.g. tweets per second, minute, hour, or day.

Interviewees in Chapter 3 also discussed reading and coding tweets as part of a qualitative or mixed methods analysis. While Agave's user interface for searching and reading tweets, and for comparing and contrasting search queries, may make it useful as an exploratory tool in a qualitative or mixed methods research project, more work is needed to investigate its applicability in this type of research. The ready access to raw tweet data afforded by Agave could make it appealing from a qualitative perspective, but additional features may be needed, such as saving and exporting selections of data, some form of coding or tagging support, and fine-grained control over which tweets are displayed (e.g. advanced filtering controls or alternative sorting criteria). Bhowmick has discussed the potential role of visual analytics tools in qualitative research (Bhowmick, 2006).

Future studies should also address how people use social media visualization systems in the context of their own projects, e.g. with their own data and colleagues. The collaborative interaction we observed in our study was limited by the unfamiliarity of the dataset and short-term involvement with the system. In a longer-term evaluation in context, it would also be possible to determine how useful Agave could be across different phases of analysis. While it is designed primarily for early-phase exploratory analysis, when researchers need to interact with an overview of the dataset in a divergent, brainstorming mode, the threaded discussions in Agave may also help researchers build and converge towards refined questions based on exploring the data. In addition, the dual filtering interface design for addressing compare and contrast questions could be useful in later phases of analysis. Understanding how technology can best support the changing practices of researchers over the lifecycle of a project is an important question for future work.

# Chapter 6: Coding Tools for Chat Data

While statistics, machine learning, and visualization can all be used to analyze social media and online communication data, many social science researchers working with these datasets also use qualitative data analysis. As discussed in Chapter 2, studying online social data from a qualitative perspective can produce a deep, rich understanding of human behaviors in context (Agarwal, Bennett, Johnson, & Walker, 2014; Aragon, Poon, Monroy-Hernández, & Aragon, 2009; De Choudhury, Counts, & Horvitz, 2013; Goggins, Laffey, & Gallagher, 2011; Starbird et al., 2015). The use of qualitative methods was discussed by the social scientists interviewed in Chapter 3, who used coding techniques to explore their datasets and to organize it into meaningful categories.

However, qualitative coding is a labor-intensive process of manually reading and interpreting large amounts of data, and it is time-consuming. For traditional qualitative data, e.g. interviews, notes, and other documents, a variety of qualitative data analysis (QDA) tools are available to help reduce the burden of "clerical" data handling tasks (St John & Johnson, 2000). Such tools make it easier to organize, code, and retrieve data, and can improve the thoroughness and auditability of qualitative analysis. However, these tools are not easily adapted to large online communication datasets, because the format and structure of the datasets differ substantially from traditional qualitative data types. Several researchers in Chapter 3 reported using general-purpose data analysis software such as Microsoft Excel or Google Spreadsheets to code Twitter data. While such tools can be effective, they are not designed for qualitative analysis and they lack many useful features to make coding easier and more effective.

Little is known about the design space for qualitative coding software. While practitioners and theorists of qualitative research have published several case studies and critiques of QDA software

(Banner & Albarrran, 2009; Humble, 2012), there has been little empirical evaluation of QDA tools, and no investigation of opportunities and design implications for coding software to analyze social media and online communication datasets.

As part of our work studying affect and emotion in the Supernova Factory collaborative chat corpus (also discussed in Chapter 4), our team undertook the task of manually coding a large quantity of text-based chat data. To help our group more easily read and explore the chat logs, I created a "chat visualization" tool. Over several months, this tool evolved into a web-based application for collaborative qualitative coding. Members of our research group used this coding tool on a weekly basis to analyze thousands of chat messages in the dataset, both for constructing a grounded taxonomy of affect (Scott et al., 2012), and for creating a sizeable ground truth training set for machine learning, discussed in Chapter 4.

This chapter presents a case study on the applied design and development of Text Prizm, our group's qualitative coding tool for chat data. Reflecting our experiences, with supplemental support from a retrospective analysis of digital collaboration artifacts such as group emails and software commit histories, I discuss the trajectory of our research as it relates to the development of the coding tool. I explain the problems we encountered, our design decisions, and our process of developing and testing prototypes at various stages of the research project. Reflecting on our experiences, I discuss implications for future work on qualitative data analysis software design and practical applications, with a special focus on tools for working with short text-based messaging data, such as social media and informal online communication.

## 6.1 Acknowledgments

The qualitative coding software discussed in this chapter was designed, built, and evaluated in 2011 and 2012 as part of a collaborative research group with major contributions from Katie Kuksenok, John J. Robinson, Daniel Perry, Taylor Jackson Scott, Ona Anicello, Megan K. Torkildson, Ariana Zukowski, Paul Harris, and Cecilia Aragon. The analysis of group processes and design decisions included here was completed later, by the author.

## 6.2 Qualitative Data Analysis Software

In this section, I discuss design issues pertaining to qualitative data analysis software, including existing software packages and literature on coding qualitative data.

As discussed in Chapter 2, qualitative data analysis (QDA) practices are typically labor intensive and time consuming, requiring extensive interpretation of relatively unstructured source materials (e.g. transcripts, notes, photos, videos). Accompanying the move to digital and electronic media over the past several decades, a variety of software tools for organizing and facilitating QDA practices have entered the market, including ATLAS.ti, NVivo, NUDIST, MaxQDA, HyperRESEARCH, and QDA Miner (Drisko, 2006). Recently, web-based tools such as Saturate and Dedoose have also become popular. These tools offer a wide range of features to help with aspects of qualitative analysis, including transcription, data collation and cleaning, organization, tracking, reading, comparison, coding, and theory building.

The breadth and complexity of options available on the market can make it difficult for qualitative researchers to choose a tool appropriate to their research project, since the tradeoffs of using various QDA software packages can be quite complex. For those deciding to use QDA software, there is a great deal of information available about best practices and risks (Humble, 2012; M.

Jones, 2007; McLafferty & Farley, 2006; Richards, 2014). While articles about QDA software currently on the market raise important issues, it is important also to keep in mind that the feature offerings of these tools continue to evolve and change, so care must be taken to re-check specific statements about their capabilities.

Aside from the benefits of reducing manual and clerical tasks, saving time, covering more data, increased flexibility, and improved validity and auditability, St John and Johnson raised concerns with QDA software, such as rigidity of research process, privileging of coding and search, focus on volume and breadth over depth and meaning, the learning curve, expense, and distraction from the focus of analysis (St John & Johnson, 2000). In cardiovascular nursing, Banner and Albarrran reflect on critical attitudes and resistance to QDA software. They point out that many QDA software packages incorporate the language and processes of specific methodologies, and that software adoption may accidentally interfere with selecting the most appropriate methodology. Additionally, they suggest that technological approaches are sometimes associated with quantitative approaches, "number crunchers", and positivist research processes, which may undermine the philosophical commitments of qualitative research (Banner & Albarrran, 2009; Coffey, Holbrook, & Atkinson, 1996).

In a similar vein, Goble et al. discuss the use of QDA software NVivo in a phenomenological research project in the field of nursing (Goble, Austin, Larsen, Kreitzer, & Brintnell, 2012), attempting to identify its influences and impact on their research process and cautioning researchers against methodological and philosophical difficulties in adopting QDA software: "[QDA] programs can impede phenomenological analysis by creating practical conditions that are markedly *unphenomenological*". On the other hand, some researchers have found the benefits of QDA software to outweigh its costs. Bringer et al. use a reflective case study approach to show

how NVivo was effective in a constructivist grounded theory project; with the help of the tool, the authors were able to move quickly back and forth between open coding and focused coding, while writing conceptual and theoretical memos (Bringer, Johnston, & Brackenridge, 2006).

Looking beyond these traditional forms of software support for QDA processes, Bhowmick has argued that visualization tools could help qualitative researchers develop stronger exploratory capabilities with their data, and better handle geographic and temporal data (Bhowmick, 2006). In addition, for online communication and social media data, many conventional tools for qualitative data analysis are not appropriate or optimal because the structure of online communication data is unlike traditional qualitative data. In the project discussed here, the scale of the data we needed to analyze, and the combination of qualitative and computational methods we wished to apply, made existing tools unusable. I will elaborate on needs and design constraints for qualitative researchers studying chat log data as I discuss the iterative design and use of our qualitative chat log coding tool in the following sections.

## 6.3 Method

The following sections discuss the timeline of our research group, our design process and decision-making, the coding software we created, and reflections on using the software. Building on Frayling's *research through design* (Frayling, 1994), Zimmerman et al. have argued that the act of designing and creating technology artifacts can constitute an important contribution to human-computer interaction research (Zimmerman, Forlizzi, & Evenson, 2007). By focusing on process, invention, relevance, and extensibility, designers can apply their talents for solving under-constrained problems and produce concrete crystallizations or integrations of knowledge that facilitate the transfer of research ideas into practice.

In this chapter, I have taken pains to carefully document our group's design process and explain the practical importance of the decisions we made. However, while the original work discussed here was conducted primarily over our group's first year of activity in 2011 and 2012, we did not plan to treat this project as a case study on the design of qualitative coding tools; as a result, the analysis for this chapter was completed later, in 2015. Because of this delay, the recollections and interpretations in the account below are less immediate. To minimize problems with completing the case study retrospectively, I gathered over 250 emails and discussion forum posts exchanged by group members during our first year, organized them chronologically, and wrote summary notes of the communications.

As primary data, I considered my notes from these emails, our shared wiki site and meeting notes, our software source control change logs, and various other artifacts (e.g. spreadsheets, presentations, wireframes, and raw data files); out of these, I pieced together the account of our activities below. Our group's collaboration style, heavily reliant on email and other digital artifacts, ensures that these resources provide a relatively comprehensive account of group activities. While these data do have gaps and limitations because they were not created with the explicit purpose of thoroughly documenting our design process, our digital communications and collaboration artifacts still help to ground my recollections and interpretations. Based on the following account of our process for designing and developing our collaborative coding software, I develop a set of challenges and implications for the design of coding tools for online social data.

## 6.4 Coding Tool Case Study

In this section, I present a narrative case study account of our group's experience with the design, implementation, and used of qualitative coding tools to analyze a large chat dataset. The period of

activity discussed here extended over three academic quarters, from September 2011 to June 2012, during which our group included from four to ten researchers.

### 6.4.1 Exploring the Chat Dataset

In September 2011, a group of colleagues and I formed a research group to study emotion in text-based communication for online collaborative work. The group's goal, as advertised to new group members, was to analyze a specific large online chat dataset that was available to us. We would attempt to detect and classify expressions of emotion and affect in this dataset, and relate emotions to events occurring in the chat log. The group planned to use "manual coding" in a grounded theory-based method, as well as automated machine learning classification methods.

During its first quarter, the research group included four junior doctoral students, including myself, and one faculty adviser. In this period, we gathered, read, and discussed articles analyzing emotion, sentiment, and affect in chat and other online communication media. These included theory-oriented articles about the social science of emotion in online communication, as well as technique-oriented articles about automatically detecting emotional expression in digital records of online communication. The group used a shared discussion forum to post ideas and questions about these readings, which we discussed in weekly meetings.

While we worked to understand the literature, group members also downloaded and began to manually explore the chat logs which were our main focus over the next few quarters. Discussed previously in Chapter 4, our corpus consisted of chat logs collected from the Nearby Supernova Factory (Aragon et al., 2009), an international astrophysics collaboration. The corpus was organized chronologically as a collection of over 1000 nightly 24-hour chat transcripts, referred to as "logs" or "sessions." These collectively contained over 400,000 chat messages from a group

chat room used by about 30 astrophysicists to discuss the operation of a telescope over a four-year period. The astrophysicists in the chat room also interact with “Bert”, an automatic chat bot that can answer simple questions for the scientists about the state of the telescope. A selection of raw (anonymized) chat logs is depicted in Figure 6.1.

```
2004-11-30T17:04:59 UTC -> Bob in scopechat: #8 is that one
2004-11-30T17:05:19 UTC -> Nathan in scopechat: bert, what is telescope?
2004-11-30T17:05:26 UTC -> Nathan in scopechat: what is the Tel Off one?
2004-11-30T17:05:31 UTC -> Bob in scopechat: Ummm
2004-11-30T17:05:40 UTC -> Bob in scopechat: bert what is te?
2004-11-30T17:05:43 UTC -> Nathan in scopechat: bert, what is tel
2004-11-30T17:05:44 UTC -> Bob in scopechat: bert what's tel?
2004-11-30T17:05:59 UTC -> Bob in scopechat: Well shucks
2004-11-30T17:06:00 UTC -> Nathan in scopechat: bert, why are you?
2004-11-30T17:06:12 UTC -> Bob in scopechat: Don't ask it that.
2004-11-30T17:06:13 UTC -> Nathan in scopechat: value_name: TEL not found in
/home/soft/snifs/Online/dev/tcs/value.map
2004-11-30T17:06:22 UTC -> Bob in scopechat: We do not want it to become sentient
2004-11-30T17:06:24 UTC -> Nathan in scopechat: I've got lots of those
2004-11-30T17:06:35 UTC -> Nathan in scopechat: I forgot to email them to you
2004-11-30T17:06:43 UTC -> Nathan in scopechat: every 30 min
2004-11-30T17:07:01 UTC -> Bob in scopechat: tom.pl did not run -- evidently no subs yet
2004-11-30T17:07:31 UTC -> Bob in scopechat: jean is it just send_SMS?
2004-11-30T17:07:40 UTC -> Bob in scopechat: So "send_SMS 'hi jean'" will work?
2004-11-30T17:07:58 UTC -> Nathan in scopechat: bert, what's active_l?
2004-11-30T17:07:58 UTC -> BERT in scopechat: [Tel Off]
2004-11-30T17:08:02 UTC -> Nathan in scopechat: there we go
2004-11-30T17:08:14 UTC -> Bob in scopechat: Why is it called that?
```

**Figure 6.1: A selection of the astrophysics chat log from October 2004. The participants are debugging a problem their software.**

Early emails within the group included questions about the naming conventions and formatting of these log files. As the group began working with the data, and reading articles on affect and emotion in text, we began to brainstorm ideas about what signals chat participants could use in their messages to convey emotion, such as punctuation, emoticons, and distorted English, as well as research questions about the relationship between group activities and emotional expression in the chat logs. We began searching for examples of emotion in the logs, and sharing our findings with the group:

> Here's what I've got so far. Not great, I think. I'm having problems finding instances of clear emotion. Because of that I just sort of annotated the log sections I thought were most emotional. *[Attached text file containing passages with emotion interpretation]*
>
> *Group member #1*

> Reply: I agree that it is challenging clarifying WHAT the emotions are. Attached is what I have so far. I'll see if I can categorize further before we meet. *[Attached spreadsheet containing annotated passages]*
>
> *Group member #2*

As these emails indicate, finding good examples of emotion in the chat logs was challenging at first. We worked on gathering information about the participants and activities present in the chat logs to better understand what was being discussed.

### 6.4.2 Initial ChatVis Prototype

We allowed several weeks for simply exploring and reading the chat log data. To help structure this process, we decided to individually collect examples of particular emotions in spreadsheets, which we shared and discussed as a group. It was during this phase that the challenges of working with the chat log data in its original form first became acute. We encountered several challenges trying to read and understand the raw chat log files (Figure 6.1). Inconsistent alignment and poor timestamp formatting created distractions from interpreting the messages. Because the chat room was persistent, and participants could join and leave, it was also difficult to keep track of which participants were present at any given moment. Finally, because the log files show all chat

messages in a uniform sequence, it was also difficult to understand the real-time pace of chat and locate the most and least active portions.

Frustrated by the difficulty of navigating through disparate nightly log files and the poor readability of the plain text chat records, I developed a simple tool to facilitate data exploration.

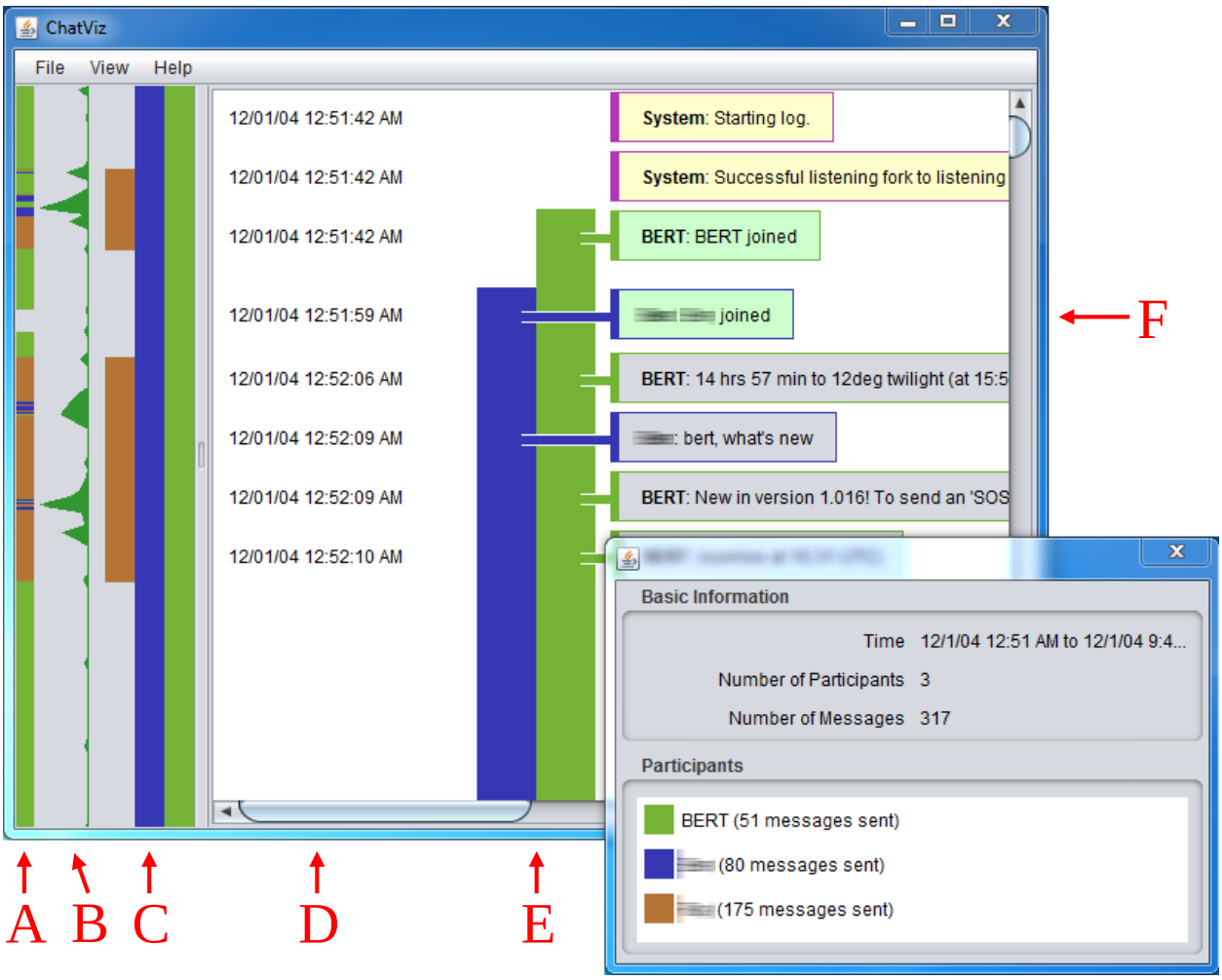


**Figure 6.2: The initial "chat visualization" prototype. Parses raw chat logs and displays the results in a readable format with metadata, including (A) who is speaking, (B) chat rate, (C) who is present in the chat room. The main panel includes (D) message timestamps, (E) participant presence bars, and (F) color-coded messages.**

This program, written in Java and initially discussed within the group as a "chat visualization," parsed the raw log files and displayed them in a user-friendly format. The initial version of this

software is illustrated in Figure 6.2. I sent the tool to the group in an email, with the following explanation:

> I was getting frustrated with how difficult it was to remember to pay attention to the timestamps in the chat logs, so I wrote a program yesterday to help visualize the conversations over time.

The three visual displays of time-oriented chat metadata in Figure 6.2 (A, B, and C) show the user an overview of the chat log that reveals who is talking at what times, when the participants were most active over the 24-hour log period, and when participants joined and departed from the room. This high-level visualization of the log was designed to help with navigating the log and with understanding the log's real-time pacing. The display of chat messages in the middle of the screen (D, E, and F) show the chat in a format similar to how the original participants would have seen it, with the addition of color-coded "presence bars" that emphasize who is currently in the chat room, and the conversational back and forth between participants. When the number of participants in the room grows larger, these bars were intended to help maintain clear awareness of the current context. Clicking on the visualizations in the left panel causes the main panel to scroll to the indicated portion of the log.

The initial ChatVis prototype was shared with the research group along with instructions for how to use it to open and view the chat log data files. After some rapid debugging, other group members were able to run the program, and reported that it would "make manually browsing the chat logs so much easier." Over the weeks following this email, I got feedback from members of the research group on the prototype, while the group continued to explore the chat data and read articles. I made

several bug fixes and small improvements to the chat visualization tool, such as adding a progress bar, flagging of problematic messages, and performance improvements.

### 6.4.3 Collaborative Coding

As the group gained increased familiarity with the chat log data and its emotional and effective content, we decided to begin coding the data with different emotions so that we could systematically enumerate and index the emotional aspects of the data. Our overall approach to this qualitative analysis was adapted from grounded theory (Charmaz, 2006). Scott et al. have extensively discussed the rationale for this methodology (Scott et al., 2012). The first phase of grounded theory is open coding, where researchers review data and create or propose codes as-needed, based on what they see in the data. In our exploration of the dataset we had noticed many instances of mixed and overlapping emotion; as a result, we often used multiple codes per message.

Our project had dual goals: both the grounded construction of a framework for affect in chat, and the creation of training data for machine learning. We wanted to be able to code our data at the message level because we felt this was the most natural way to organize the data for machine learning. With whatever tool we used, we also needed to be able to recover structured coded data from the software. We examined existing QDA software to determine whether it would meet our needs. While most QDA software supports open coding, we could not find any tools that would work with logs of thousands of very short messages, coded at the message level. Traditional QDA is conducted on long documents such as interviews and notes, where coding is applied at the level of episodes or passages (e.g. sentences, paragraphs). While we could load a large chat log into such software, it would not naturally structure the data in such a way that we could easily recover the coded data at the level of individual messages.

Perhaps the most appropriate tools available were standard spreadsheet programs such as Microsoft Excel or Google Spreadsheets. These would have presented the data in a reasonably readable way, and would have permitted coding at the message level. Spreadsheets are also relatively easily converted to machine-readable training data for machine learning algorithms. However, spreadsheets also had some problems. Their tabular representation was awkward for the style of coding we wanted to do, with multiple codes potentially overlapping on individual messages. In addition, collaboration with spreadsheets would require copying, sharing, and reintegrating the coding in many spreadsheet files; this becomes cumbersome for large datasets.

In summary, existing tools either did not support the process we wanted to use, or did so in an awkward way that would have limited our efficiency. Because we already had the ChatVis software, which provided a usable way of navigating and reading the chat log data, we decided to add coding functionality to this tool. A very simple server-side PHP program was created, which presented a "code synchronization" API over HTTP. ChatVis client programs could save and retrieve codes using this API, and the PHP program persisted these codes to a MySQL database. Once these extensions were working, in mid-November 2011, we developed a plan for how to start coding the data. We decided to code for moods, mood transition, affect on specific messages, and any other phenomena related to emotion that we thought were interesting. I provided instructions for how to run the updated software, how to connect it to the code persistence server, and how to interact with it to code the chat logs.

While the group tried the tool out over the next couple of weeks, I continued adding new features and improvements, including a search box, a feature for marking important messages with a "star", and customization of the way "presence bars" were displayed. We also refined the visualizations in the left-hand panel of ChatVis, presenting metadata about the presence of chat participants in a

more compact format (Figure 6.3). Our selection of data for coding in these early phases was purposeful, guided by trying to find the most significant and theoretically rich chat logs:

> Let's try to find some log files that definitely have some group-work/collaboration that results in the completion of some objective […] I think that as a next step, manually coding some logs that definitely have emotional content surrounding some event would be extremely beneficial.
>
> *Group member*

As the number of affect and emotion codes created by group members during open coding continued to increase, the flat list that the tool used to display codes became unwieldy. We switched to a hierarchical coding system, and added the ability for the tool to display and select codes in a hierarchical fashion. At the end of the first quarter (December), the group also ran an initial experiment on using machine learning techniques to analyze the chat data, and began learning Weka, a popular open-source machine learning toolkit in Java. We also added the ability to export coded data from ChatVis to Weka's ARFF format. However, after this, no other significant features or enhancements were added to this Java-based version of the ChatVis software. A screenshot of the redesigned ChatVis prototype is in Figure 6.3.

At the end of our first quarter as a research group, we had used the collaborative coding features of ChatVis to code about 1000 chat messages. While the number of messages coded was not large at this point, we had created 68 distinct affect codes; group members had their own preferred subsets of codes, and the distinctions between them were nuanced. We began discussing how to bring a more manageable structure to our code scheme. In an email, I wrote to the group:

> We need to talk about our coding strategy in our meeting tomorrow. I think we're all taking very different approaches to coding at this point.

The section below discusses the process of consolidating our codes to a more focused framework.

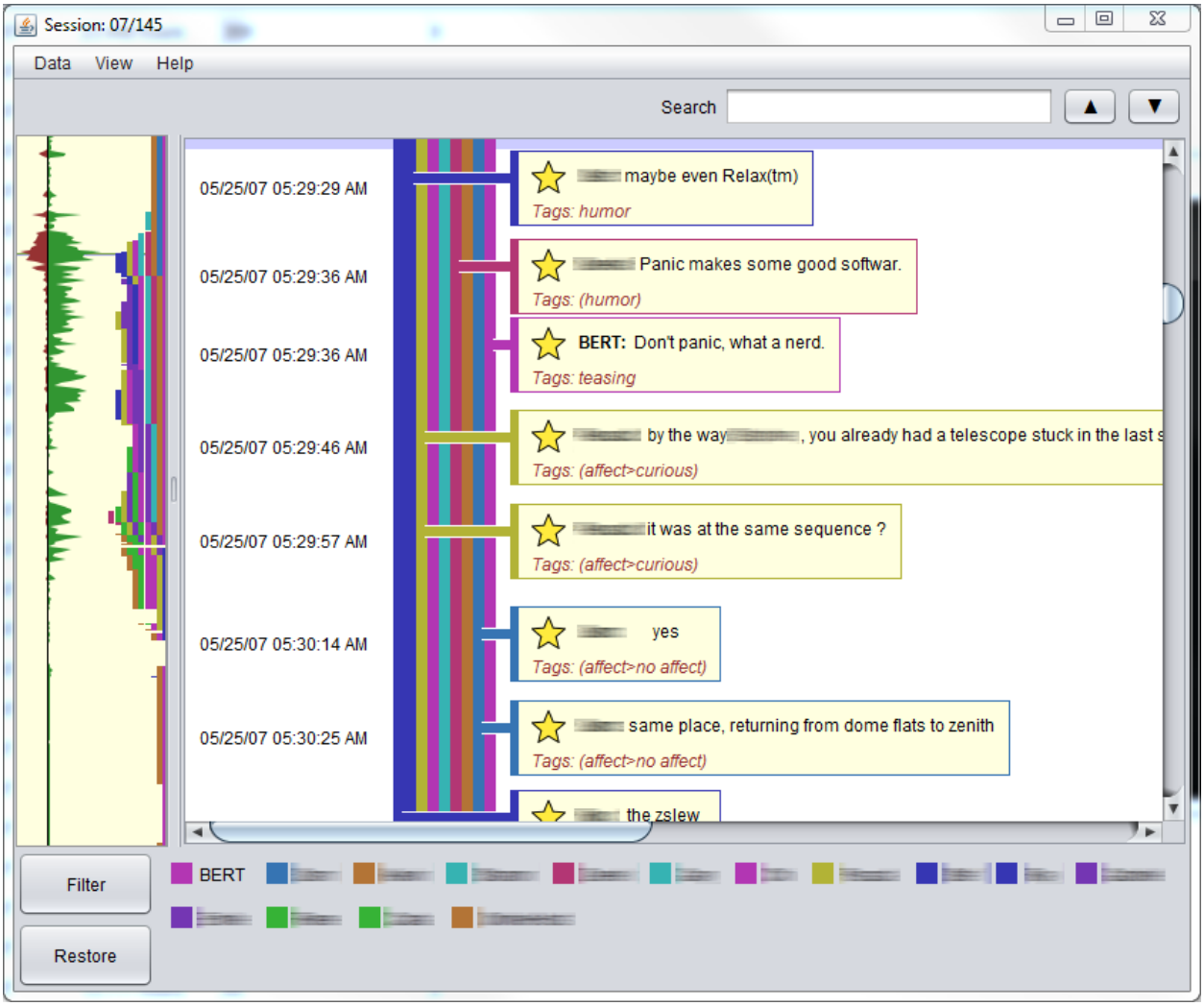


**Figure 6.3: Final iteration of the Java-based ChatVis coding tool. Provides search and filtering, and visualizations in the left panel. Codes are applied by clicking on messages, and displayed under each message. Codes are persisted to a database server.**

### 6.4.4 Consolidation of Coding Scheme

Having used ChatVis for several weeks at the end of our first quarter to create a large number of codes grounded in the chat data, the focus of the group turned, in our second quarter, towards

refining our coding scheme and unifying our coding strategies. Throughout this phase, we continued coding; by late January, we had collaboratively coded messages from six different logs. As a group, we discussed each code and identified similar codes that could be combined, codes that could be removed, and ways to organize the codes better (Figure 6.4). Addition and modification of codes was accomplished "manually," through back-end database queries.

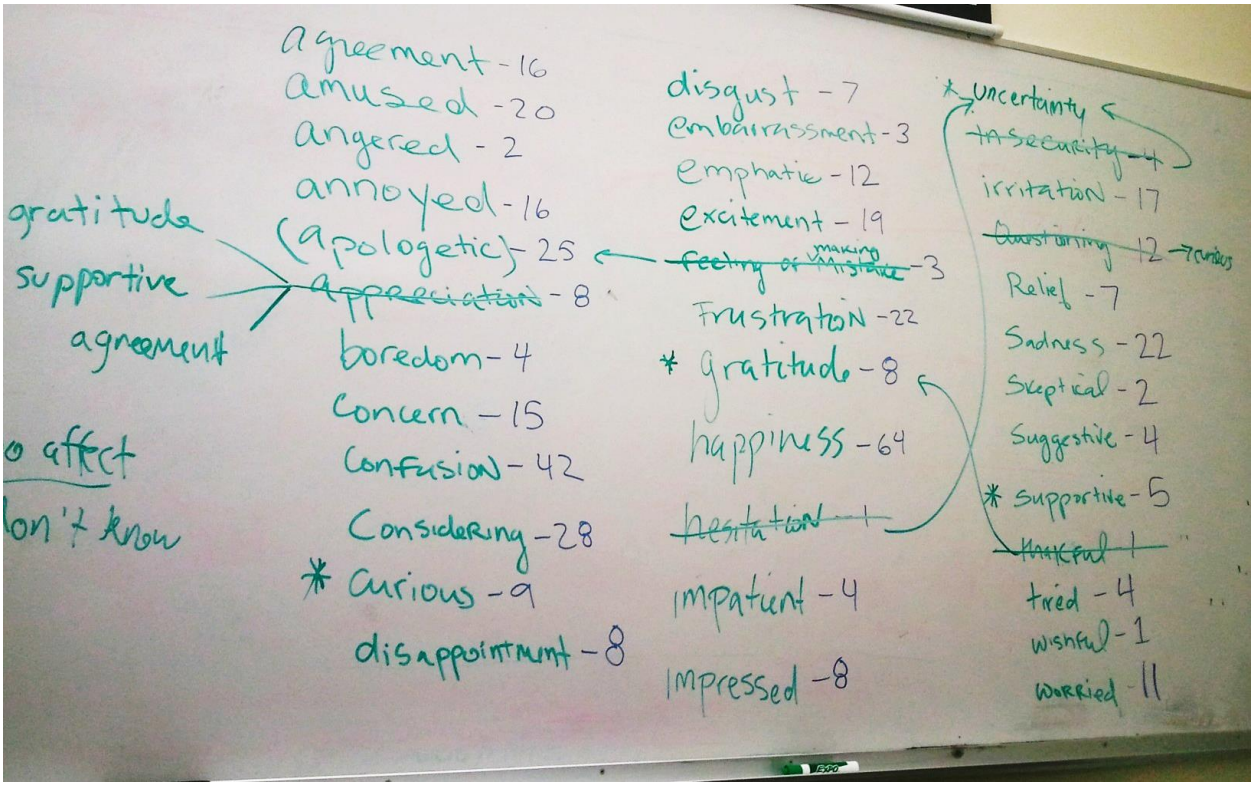


**Figure 6.4: A whiteboard discussion refining our coding scheme. The numbers beside each code refer to frequencies of use in the coded data.**

We held extensive group discussions to resolve ambiguities in our interpretations, e.g. Figure 6.4. With a large set of codes, many of which appeared closely related, it was often unclear which codes to use on particular messages. We began working on a "code book" of definitions and examples for our codes. From the literature on emotion in text communication, we also incorporated concepts from existing theories of emotion. Specifically, we compared our codes to

the system of emotions developed by Robert Plutchik (Plutchik, 2001), and incorporated some of these emotion categories into our code scheme (Figure 6.5).

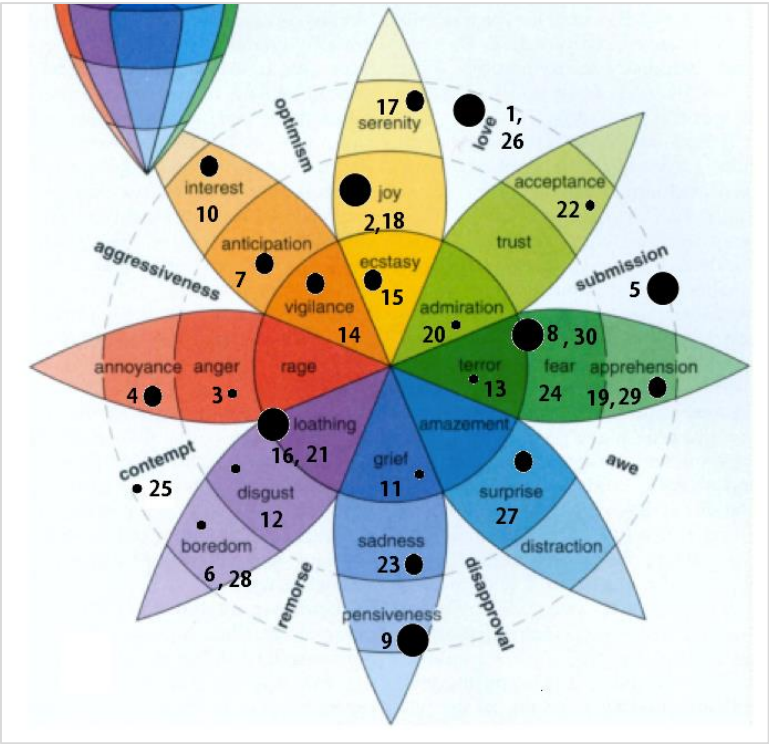


**Figure 6.5: Mapping our affect codes onto the Plutchik "wheel" of emotions. Our codes are indicated by numbered black dots.**

Our group's second goal was to apply machine learning to our dataset. Two new students with computer science backgrounds joined the group, in part to help develop software and run machine learning experiments on the data. While we continued to discuss and develop our affect code scheme, we also imported the chat data into MySQL, began running exploratory quantitative analyses, and created visualizations of the coded data in Microsoft Excel and Tableau; this process echoes that of the social media researchers interviewed in Chapter 3. We analyzed the frequencies of different words and other text patterns in the data, created visualizations comparing codes to different textual features, and brainstormed other analyses we could carry out.

After several weeks of exploratory analysis and iteration on our affect codes, we had refined our code scheme and created a code book containing 30 distinct categories of affect with definitions and examples. At this point, we had applied roughly 4,200 codes over about 3,800 chat messages in the Java-based version of ChatVis, but this coding had been done with our older, messier code scheme. For the next phase of the project, we abandoned this work as preliminary coding, and started fresh with our smaller, focused coding scheme. Satisfied with the descriptive power and efficiency of our coding scheme, we determined that our primary goal for the final quarter of the project would be to broaden the scope of our coding to a much larger amount of chat data. This required a re-examination of ChatVis, and eventually led to its redesign, discussed below.

### 6.4.5 Critical Design Reflections

While the initial ChatVis tool was successful in helping our group explore, read, and collaboratively code the chat data, it had a number of limitations. The group met to reflect on the ChatVis tool and discuss what worked well, and what needed improvement. Because our needs were shifting towards coding large amounts of chat data, we also reconsidered several competing QDA tools, including NVivo, ELAN, ATLAS.ti, Saturate, and PCAT (now DiscoverText).

The group determined that ChatVis was simple to use and reliably performed the collaborative coding function, but that the selection and deselection of codes was difficult, the timeline was a bit confusing and unclear, and keyboard-based interaction was lacking. Before attempting to code large amounts of data, the group identified desirable changes, including:

- Easier coding interaction, including keyboard coding.
- A personal view of codes (i.e. hide other people's coding).
- A progress tracker that promotes a sense of accomplishment.

- Incorporation of code definitions and examples into the UI.

In order to incorporate improvements and facilitate coding more data than we had managed to code in the past, we decided to redesign and re-implement the coding tool as a web application. Because of ChatVis's humble beginnings as a tool intended simply to make the raw log files readable, the architecture of the system presented difficulties for further refinement and additions. Rebuilding the coding tool as a web application offered easier deployment and testing, which was ideal for the rapid progress we wanted to make over the coming quarter. Because users could simply follow a link in their browsers to open the application, the web provides an easier user experience than a Java application. Moreover, we decided that HTML, CSS, and JavaScript would provide a more expressive medium for implementing visualizations and interacting with the data.

### 6.4.6 Redesign and Transition to the Web

Over several weeks, several members of our research group collaborated closely to implement our new web-based coding tool, which we called "Text Prizm" after its colorful user interface. With a better understanding of the workflow we were now designing for, our research group held several design brainstorming sessions to shape the next version of the tool. We implemented the coding tool as a PHP web application with CodeIgniter (a popular PHP web application framework), using a MySQL database back-end. Our front-end application was implemented with Backbone.js (offering elements of a JavaScript model-view-controller framework), jQuery, and D3. Early versions of the application included the following general capabilities:

- Loading and viewing chat log data in a layout similar to the earlier ChatVis program.
- Collaboratively applying codes to the chat data.
- Searching the logs, and finding examples of messages with specific codes.

- A chronological visualization providing an index into the chat log data for coding.

A screenshot of the coding portion of Text Prizm is provided in Figure 4.1. Messages are loaded and displayed in the middle portion of the screen, while codes are displayed on the right. The URL used to access the coding tool encodes which portion of the chat data is being coded, making it easy for group members to store and share references to the data. The codes are displayed in a compact and readable format as part of each message, and only the codes for the current user are displayed. We show a progress bar at the top of the UI indicating the percent of messages in this log that have been coded. Codes can be created using the box in the top right of the screen.

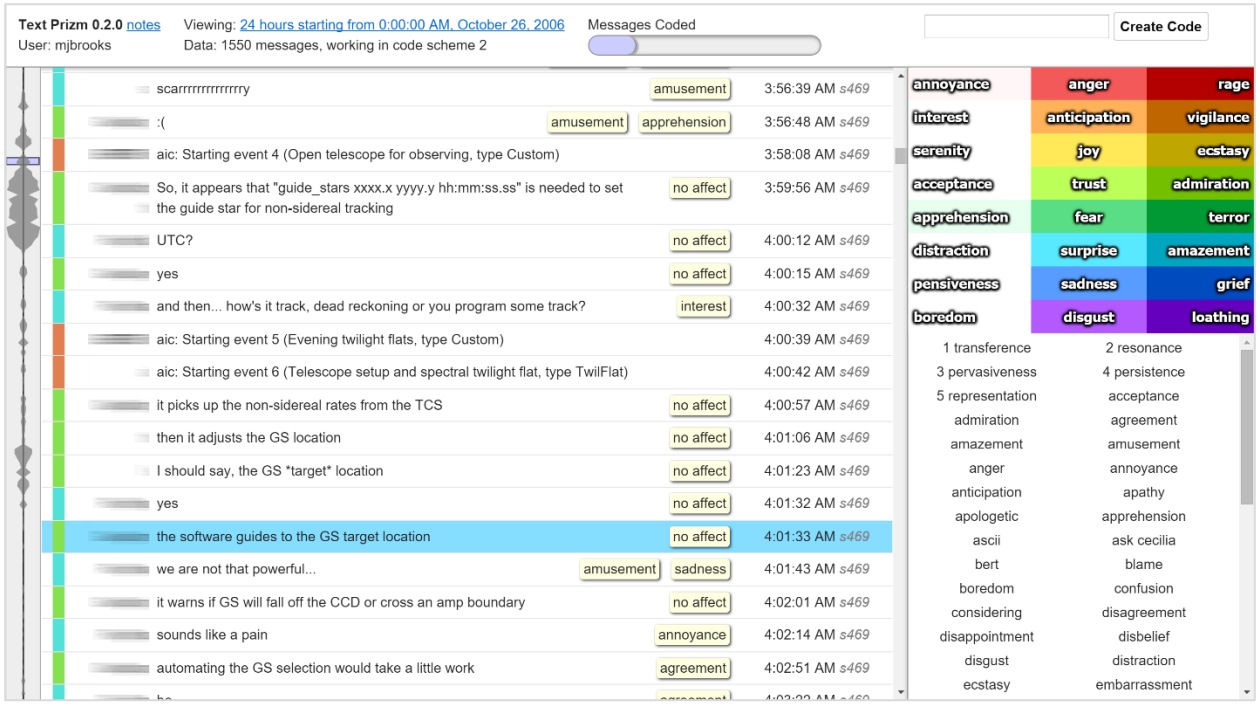


**Figure 6.6. The web-based coding tool Text Prizm. The visualization on the left depicts message rate over the 24-hour log. Color bars on the left of each chat message indicate the chat participant. Codes are available in the right panel, or through keyboard interaction.**

We preserved some of the "visualization" features from the original Java-based ChatVis prototype, but simplified the design by removing the display of participant presence. As in the earlier version, a vertical strip along the left of the screen presents the rate of chat messages over time within the

24-hour period of the current log. As the user scrolls up and down in the log, a corresponding indicator moves up and down along the visualization to provide context for the current view. The visualization can also be clicked to jump to another point in the chat log.

In the web-based coding tool, we focused on creating a highly efficient coding interaction to enable users to code a large number of chat messages quickly. Rather than using a context-menu to select codes as in the original prototype, here we display the codes at all times, in a sidebar on the right of the screen. However, with approximately thirty different codes available, it is challenging to display the codes in a way that helps the user select codes in an unbiased way. We were concerned that coders might work with only a handful of frequently-used familiar codes, or only those at the top of the list. To help ensure a well-rounded representation of emotions in the log, we displayed the codes in two ways. The lower-right panel shows a complete list all of the available codes and requires scrolling to bring all of them on screen. The top, "rainbow" color-coded portion pulls out the Plutchik emotion codes and displays them in a grid that emphasizes the full spectrum of emotional "hues" and "intensities."

Based on our group's experience using the early prototype, we had learned that clicking and selecting codes quickly becomes a frustratingly inefficient way of coding chat data. Keyboard-based interactions can be efficient, but are generally thought of as expert-oriented features because they are difficult to discover and must be memorized, while graphical menus and buttons are accessible for beginners (Lane, Napier, Peres, & Sandor, 2005; Tak, Westendorp, & van Rooij, 2013). Given that we expected users to be processing several thousand lines of chat, we decided that the efficiency offered by keyboard shortcuts would be appreciated, particularly if we could design a fairly natural mapping of keystrokes to codes.

To help make coding efficient, we developed the capability of coding using solely keyboard interactions. First, the user is able to move a "message selector" (blue box in Figure 4.1) through the chat log using arrow keys. Next, with a message selected, the user may press a key to bring up a dialog offering keyboard-based code auto-completion (Figure 6.7). The "enter" key causes the code to be applied. With only a few key presses, any of the 30 affect codes can be selected.

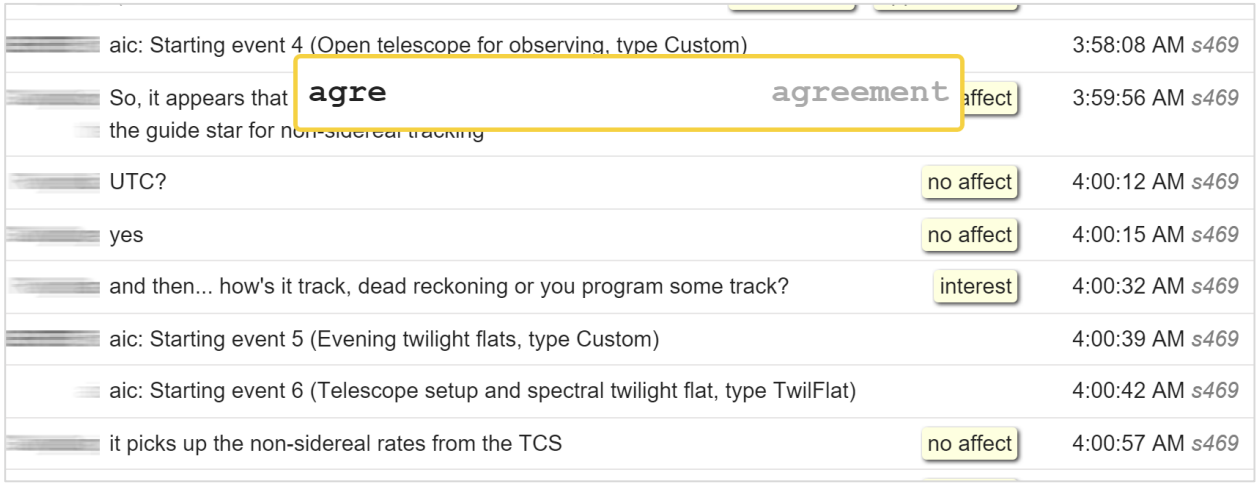


**Figure 6.7: Keyboard-based selection of codes. When a chat message is selected and the user begins typing, a code selector appears that will provide automatic code completion.**

After several weeks of coding with the new tool, we decided that, in addition to our many affect codes, we would begin to code the chat messages for "valence," or the positive/negative aspect of emotion, and "intensity," the level of emotional arousal. To make these new codes easier to remember and apply, we developed a specialized coding interface, shown in Figure 6.8. Pressing the slash key causes the Intensity and Valence dialog to appear, which allows selecting a valence and intensity combination using the arrow keys, WASD keys, or IJKL keys.

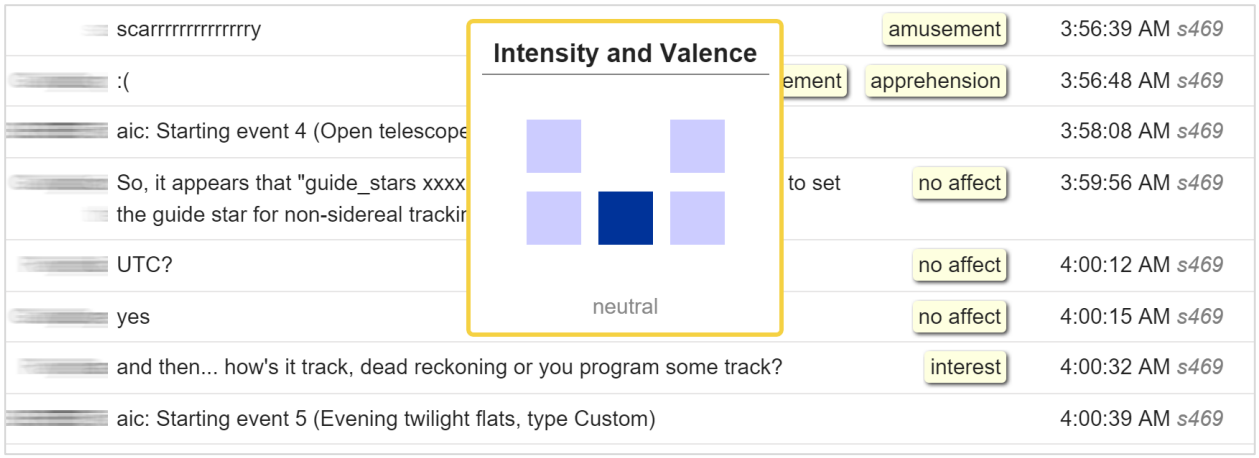


**Figure 6.8: Specialized UI for coding valence and intensity. Pressing the slash key activates a coding dialog for affect valence (negative/left, neutral/middle, positive/right) and intensity (high/up, low/down).**

In addition to the main coding portion of the web application, we also developed several supporting tools that provide additional functionality. We found over the first few months of coding that we exchanged many emails with each other with the goal of checking "is someone else coding log X" or finding out what logs have been coded. To meet this need for contextual awareness (Mark, Fuchs, & Sohlenkamp, 1997), the "Coding Stats" page displays a visualization of the chat dataset on a nightly basis from its beginning in 2004 to its end in 2008 (Figure 6.9). It shows the number of messages on each nightly log, as well as the number of messages that have been coded. Most importantly, by clicking on a square, the user can quickly enter the coding tool to work on coding the chat data. This tool was our primary portal for jumping into and coding the chat data, since it allowed us to see the chronological distribution of our work.

In addition, the "Code Browser" tool (Figure 6.10) was created to allow access to example chat data for each code. The table shows the number of times each code has been used, and the links at the right take the user to a list of examples. We frequently used this tool in group discussions to clarify ambiguities in the coding scheme and determine what the various codes mean.

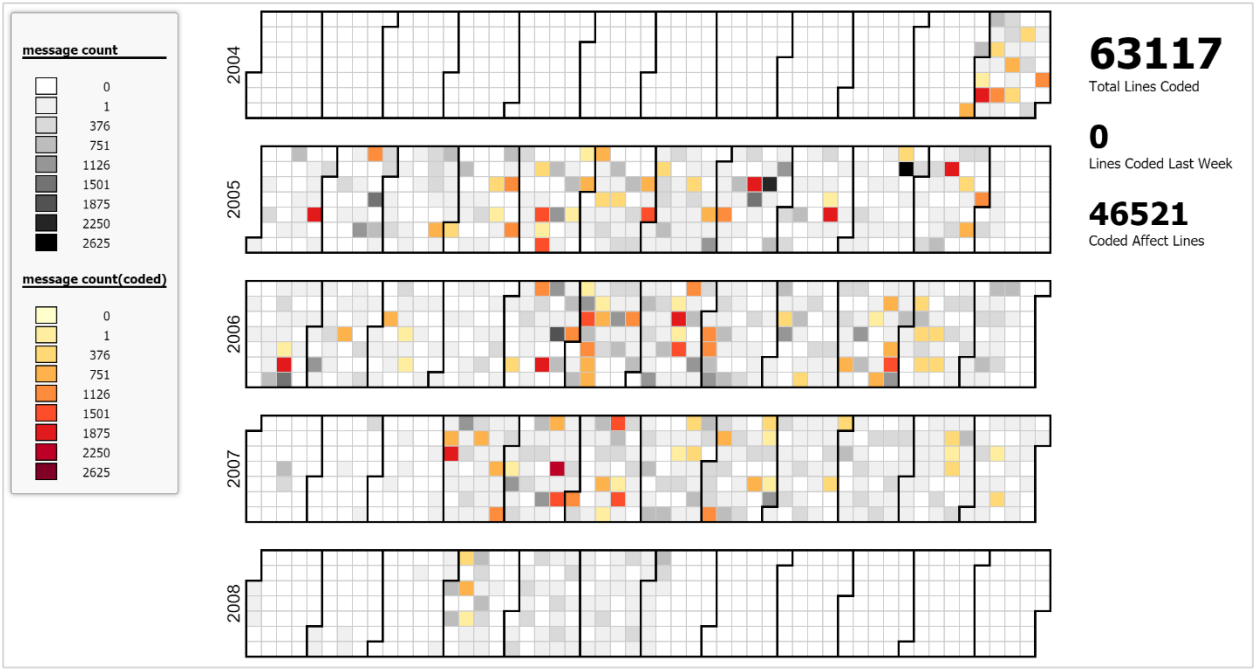


**Figure 6.9: The "Coding Stats" visualization for finding un-coded data. Provides a chronological visualization of the chat data indicating how many messages exist on each night, and how many have been coded. Clicking a square opens the coding tool.**

Search:

| Code | Flagged, Unedited by Me | Flagged, Unedited by Anyone | Flagged | Total | View Examples |
|---|---|---|---|---|---|
| 1 transference | 0 | 0 | 0 | 3 | best \| rand \| flag |
| 2 resonance | 0 | 0 | 0 | 0 | best \| rand \| flag |
| 3 pervasiveness | 0 | 0 | 0 | 0 | best \| rand \| flag |
| 4 persistence | 0 | 0 | 0 | 0 | best \| rand \| flag |
| 5 representation | 0 | 0 | 0 | 0 | best \| rand \| flag |
| acceptance | 4 | 4 | 4 | 1900 | best \| rand \| flag |
| admiration | 1 | 1 | 1 | 134 | best \| rand \| flag |
| agreement | 5 | 1 | 5 | 2729 | best \| rand \| flag |
| amazement | 0 | 0 | 0 | 91 | best \| rand \| flag |
| amusement | 6 | 5 | 6 | 4088 | best \| rand \| flag |
| anger | 7 | 0 | 7 | 137 | best \| rand \| flag |
| annoyance | 7 | 2 | 7 | 1623 | best \| rand \| flag |
| anticipation | 2 | 2 | 2 | 879 | best \| rand \| flag |
| apathy | 0 | 0 | 0 | 39 | best \| rand \| flag |
| apologetic | 1 | 1 | 1 | 326 | best \| rand \| flag |
| apprehension | 3 | 3 | 3 | 1206 | best \| rand \| flag |

**Figure 6.10: The "Code Browser" gives access to examples of each code. Also shows flagged and "high quality" examples.**

### 6.4.7 Intensive Coding

At the beginning of the final quarter of our project, in April 2012, we welcomed three new members to our research group with the goal of increasing our ability to code large quantities of chat data. With the first working prototype of Text Prizm ready for use, we provided our new members with instructions, a copy of the code book, and access to the coding tool. To help the new members acquire a common understanding of the codes, we organized group "co-working" sessions, where multiple group members would gather in the same physical location to code data together. This proved effective both for maintaining motivation and interest in coding, and for answering questions about the data and coding system.

Drawing on our extensive in-depth experience reading, interpreting, and coding the chat data, we began analyzing the language used in the chat messages, developing feature extraction techniques, and testing various machine learning algorithms. During this time, our focus was divided between conducting machine learning experiments (discussed in Chapter 4), continuing to develop and improve Text Prizm, and coding as many messages as possible.

The Text Prizm tool supported efficient coding. While our work with ChatVis over the previous months had produced about 4,200 code applications total, in the first 1.5-hour co-working session each of the four users managed to code about 450 messages (1,800 codes applied in total). After the first coding session, I sent the following progress report to the group:

> I'm glad to hear that the coding session yesterday went smoothly. […] From examining the database, it looks like each person coded about 450 chat messages during the session, which is great progress! About half of those messages were coded with emotion by at least one person. On [those messages], I checked for

agreement between the four coders on whether emotion was present. It looks like all four of you agreed there was emotion present on 85% of these messages, with the other 15% getting agreement from three people or less. That looks pretty good to me – I didn't check for agreement on specific emotions.

We'll be working on getting better progress tracking and work management systems in place so that everyone can see how they're doing in the next couple of weeks. If anyone has any ideas for features or improvements, be sure to make a note and bring it up during the group meeting.

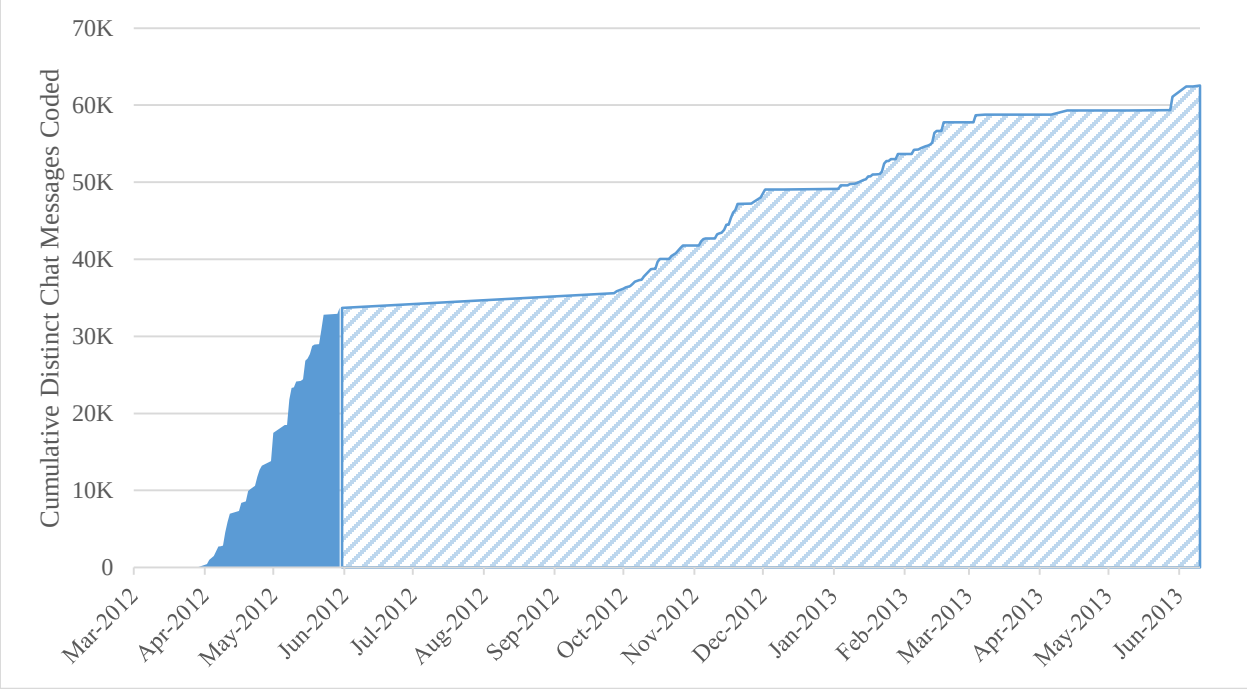


**Figure 6.11: Cumulative messages coded over time using Text Prizm. The filled portion in Spring 2012 represents the period discussed in this chapter.**

We held around 15 co-working sessions to code the data over the course of the quarter. Our core group of coders averaged about 450 messages per session at the beginning of the project, but reached an average of 1,000 messages per session by the end. The maximum number of messages coded in a session was 2,700. A chart of the cumulative number of messages coded over time is

provided in Figure 6.11; in April and May, we coded over 30,000 chat messages. Although it is not discussed here, in later work we eventually coded about 60,000 lines of chat.

Of course, quality of coding is also important. Although we had discussed inter-coder reliability and agreement in the past, the question of how to calculate the reliability of our coding once again arose. It quickly became apparent that for data coded as ours was, with a variable number of coders, a large number of codes, and multiple codes allowed for a message, we could not use commonly used metrics such as Cohen's *kappa* (Cohen, 1960) or Krippendorff's *alpha* (Hayes & Krippendorff, 2007). In order to have some quantitative sense of our reliability, we developed a simulation-based variation on *kappa* that worked for our data discussed in detail in Chapter 4. While we did not integrate this calculation into our tool at the time, having reliability and agreement statistics within the tool would have helped us monitor coding quality over time.

In our meetings, we continued to elicit feedback on the coding tool and understand what was working well and what could be improved. While using Text Prizm intensively, we added a tool for doing advanced searches over the chat dataset, as well as the valence and intensity coding interface (Figure 6.8). One group member sent the following comment in an email:

> This tool is awesome, and I'm very impressed with the hard work everyone has put into it to produce a functional and usable tool for coding chat logs… something that I don't think exists.
>
> *Group member*

As the quarter progressed, there were requests for ways to get additional information out of the system, such as visualizations and summary statistics about the coding progress, the reliability and agreement for specific codes, and the distribution of codes over the dataset. We added a set of

basic statistical tables and charts to the Coding Stats page, but there is potential for much more. Group members asked not only for summary statistics about their coding, but sophisticated ways to visually explore and navigate the dataset itself, along with ways to analyze the dataset in conjunction with their coding. We also discussed ideas for how to visually monitor collaborative coding projects like ours, in order to track progress, consistency, agreement. Despite our improvements to how codes were displayed, group members reported struggling to remember about the existence of all of the codes, and tended to rely on a smaller subset for most cases rather than the full range of 30 types of affect. Designing to mitigate memory and bias issues in a tool for high-efficiency coding is a challenge for future work. These issues are discussed in more detail in the following sections.

After May 2012, the focus of the group shifted to other projects, but we have continued to work on improvements to Text Prizm. It has been our goal to convert Text Prizm from a specialized tool custom-built and optimized for the Supernova Factory chat data and our affect coding scheme, into a general-purpose tool that could be applied to collaboratively code any similar type of dataset. Barriers to general use include the deployment and setup process, getting data ingested, and exporting coded data to other formats.

## 6.5 Challenges and Implications for Design

In this section, I will draw on the case study presented above to discuss several design challenges and constraints for systems that support collaborative coding of short text-based communication datasets, such as chat and social media.

### 6.5.1 Data handling

In large online communication and social media datasets, qualitative coding can only be applied to a small subset of the data. As discussed in Chapter 3, researchers spend a great deal of time whittling down the sprawling "raw" datasets obtained from online sources to subsets that are both relevant to their research questions, and manageable for qualitative analysis. In our project, we selected which subsets of data we would code by sampling 24-hour periods which featured greater-than-normal levels of activity, on the assumption that these active periods would include examples of affective communication and provide greater theoretical richness for developing our taxonomy of affect. Researchers may apply a wide variety of techniques and approaches for subsetting and sampling their data prior to coding. How to support this crucial part of the process in data analysis tools is an open design challenge for future work.

Researchers studying online social datasets ask a huge variety of questions and there is no way to anticipate the kinds of analysis that researchers will need to do. In our experience with ChatVis and Text Prizm, we repeatedly needed to export coded data from our tools into CSV and JSON files, SQL databases, and other formats, so that we could run analyses in Tableau, Python, Weka, and other environments. Text Prizm's use of a standard MySQL backend made it easy to restructure our data as we wished. Of course, for such access to the backend storage system to be useful, the database structure must be clearly documented. Lack of transparent, usable access to structured coded data in existing tools was a motivating factor in developing our own coding tool.

### 6.5.2 Bias in selecting codes

For the majority of the coding in our project, we worked with a large set of about 30 codes, and our preliminary coding used an even larger 60-code scheme. It is very challenging for a human

coder to give fair consideration to so many codes when evaluating a line of chat. Rather than selecting the "best" code or codes for the chat message, the coder is subject to many biases because of the limitations of memory. For example, if the most recently-coded group of chat messages focused on a small set of codes, the coder will more easily recall these codes in future messages, and in general the coder is subject to serial position effects such as "recency" and "primacy" (Murdock Jr., 1962). The user interface of the qualitative coding tool plays a role in mediating biases in the selection of codes. For example, if codes are displayed in a list, as in the Text Prizm user interface (Figure 4.1), the user is more likely to read and more likely to recall the items at the top of the list. Our use of the Plutchik emotions to structure how Text Prizm displays our codes was intended to combat this source of bias.

More research is needed to find ways to help coders make good judgments without sacrificing efficiency. Some qualitative researchers have published about the use of qualitative data analysis software, and how that software might influence their practices. Such work includes systematic reviews to assess the prevalence and adoption of various QDA software packages (Humble, 2012), the advantages and disadvantages of using QDA software (Banner & Albarrran, 2009; M. Jones, 2007; McLafferty & Farley, 2006; St John & Johnson, 2000), and case studies analyzing specific QDA packages (Bringer et al., 2006; Goble et al., 2012).

While the arguments surrounding QDA software are thoughtful and detailed, there is a lack of empirical work on how specific aspects of QDA software affect research practices. For example, it is not known how the design of a user interface for coding might affect the quality or consistency of coded data. For empirically studying how interface design relates to coding and labeling tasks, some inspiration may be taken from related work in the crowdsourcing, human-computation, and human-computer interaction communities (Mason & Suri, 2011; Quinn & Bederson, 2011). While

much of this research focuses on techniques for identifying and compensating for biased and poorly-performing crowdworkers (Tarasov, Delany, & Cullen, 2010), some researchers have studied how user interface and task design choices affect worker behavior and outcomes (Heer & Bostock, 2010; Kittur, Chi, & Suh, 2008; Toomim, Kriplean, Pörtner, & Landay, 2011). Building on these experiments, future work could compare different aspects of QDA tool design by deploying coding tasks using different interface designs on a crowdsourcing platform, such as Amazon's Mechanical Turk, and comparing the reliability and distribution of results obtained. Not all qualitative approaches stress reliability and systematic coding, but understanding how QDA software affects outcomes would lead to better-designed tools for all styles of qualitative research.

### 6.5.3 Coding quality controls

There are many ways that qualitative coding can go awry, especially when done collaboratively, over a large dataset, and over a long period of time. A classic challenge for qualitative coding practitioners is coder disagreement; that is, when coders have different interpretations that lead to inconsistencies in coding. Tools can help manage disagreement by providing various agreement and reliability statistics (Cohen, 1960; Hayes & Krippendorff, 2007). However, there are more complex dimensions to this problem. In long-term projects, understandings of codes may shift, a phenomenon sometimes called concept drift or evolution (Kulesza, Amershi, Caruana, Fisher, & Charles, 2014). Qualitative researchers also report "overcoding," when researchers become fatigued and cease to be discriminating in which codes they apply (McLafferty & Farley, 2006).

Researchers have different needs for quality control, and their needs vary over the course of a project. However, researchers using coding software should be able to understand how multiple coders differ in their interpretations, observe how those interpretations may be evolving over time,

and notice when quality suffers, e.g. from fatigue or unclear code definitions. Interactive visualizations and statistics about coding work may help researchers reflect, understand, and take control of their qualitative coding processes.

### 6.5.4 Efficient coding interactions

During the design process discussed in this chapter, we attempted to maximize the efficiency of applying codes in Text Prizm. This was accomplished through support for entirely keyboard-based interaction with the system. While our group members coded using both keyboard and mouse interactions, our highest-volume coders praised the keyboard interaction as a major time saver. While many qualitative coding tools provide keyboard shortcuts for applying codes, we stress the importance of such techniques for coding larger online communication and social media datasets. In such cases, we found that requiring the user to interrupt their coding with unnecessary mouse movements leads to significant frustration. Keyboard shortcuts are less discoverable and require memorization (Lane et al., 2005; Tak et al., 2013), so designers must take care to ensure that beginners can easily make the transition from graphical coding interactions to the keyboard.

### 6.5.5 Context support for coding

Researchers interviewed in Chapter 3 spoke of the importance of understanding context for accurately interpreting the meaning of tweets. Similarly, chat messages are usually extremely short, often with only a few words per message, and depend greatly on context. As with Twitter, the users who author and send chat messages possess a great deal of implicit contextual information that allows them to make sense of messages. For a researchers attempting to understand what chat participants are talking about, gathering context can be a major challenge and source of inefficiency.

Context in computer-mediated communication is a complex issue (R. H. Jones, 2004), but some simple design choices can affect how readily contextual information is available to coders. For example, in our tools, we decided to present chat messages in a linear stream; this way, when a researcher reads the log in order, he or she can build up an understanding of the conversational thread that mimics that of the original chat participant, and goes a long way towards clarifying contextual information. Text Prizm also included time-series visualizations that help the coder see messages within the larger chat stream, as well as visual indicators of which participants were logged into the chat room. In contrast, in the coding software DiscoverText, texts are typically presented in random order and without contextual information. In some projects, this decontextualized coding process might be desirable, in that it promotes independent, repeatable coding: it encourages the coder to think only about what is visible in the text of the message. The appropriate presentation of context will vary from project to project.

Beyond these simple forms of context support, software for coding could automatically obtain and surface outside contextual information to assist with coding. For example, for tweets, a coding tool could automatically look up summary information about Twitter user accounts and hashtags, which could be presented to the coder for context. Such techniques have been explored in visualization tools for social media data (Mazumdar, Ciravegna, Gentile, & Lanfranchi, 2012). Sophisticated computational techniques have also been developed for automatically assembling or disentangling contextual information in computer-mediated communication (Ranganathan, Campbell, Ravi, & Mahajan, 2002; Wang & Oard, 2009).

### 6.5.6 Maintaining mutual awareness

Collaboratively coding a large online communication dataset is a complex task that requires several kinds of coordination work. In the original ChatVis tool, we initially provided no means of finding out what chat logs other people were working on, or what logs had been selected for coding that week. We later added a way to see which logs had been coded, and in Text Prizm we added the Coding Stats visualization to show the distribution of coding effort (Figure 6.9), but still we exchanged many emails to find out who was working on what. In collaborative work and shared projects, it is critical to have awareness of group and individual activities (Dourish & Bellotti, 1992), and tools for collaborative coding should provide support to help coders understand what other people are working on. Awareness support can also facilitate the formation of shared conventions (Mark et al., 1997); providing information about *how* other coders are working promotes common understanding and reliable interpretation of the data and codes. In our project, group coding sessions provided this awareness and shared understanding of coding conventions.

### 6.5.7 Coding schemes and structures

In our modified grounded theory methodology (Scott et al., 2012), we developed a taxonomy of affect through multiple iterations, progressing from unstructured, open coding to focused coding. Throughout this process, we made modifications to our coding scheme as our understanding changed. The simple Code Browser tool we created for finding examples of specific codes was a useful way of reviewing what codes meant, and how they were used in practice (Figure 6.10). However, aside from adding new codes, our tools provided no mechanisms for other restructuring operations (e.g. delete, merge, split). Although most existing QDA tools do support these operations, some qualitative researchers report finding these discrete transformations of the coding scheme to be overly "definitive," unsubtle, and lacking in nuance (Goble et al., 2012). Tools may

be able to provide deeper support for refining and developing categories through introspection and diagnostic information that helps users understand which codes they may wish to modify, and the impact and meaning changing the codes in light of how they have been used in the data.

The coding schemes used in qualitative research can be structured and applied in different ways. In our project, we began with a flat code scheme, but as the number of codes grew, we sorted the codes into groups, as is commonly done in grounded theory. With this shift, we modified our software to use a hierarchical code selection menu reflecting the structure of our scheme. Later, when we reorganized and focused our code scheme to the smaller 30-code set, we returned to a flat structure, which we displayed in Text Prizm as a list. QDA tools may need to be able to adapt to different types of code scheme structures, and structures may change over the course of a project.

Codes can be applied in many different ways. It is typical in traditional qualitative analysis for the researcher to attach codes to passages of arbitrary length based on the content and the code, such as sentences, paragraphs, or even individual words. We coded our data at the chat message level, but at various times we considered using sub-message coding to focus on textual features for communicating emotion; at other times, we considered coding groups of messages, e.g. conversations. In Chapter 3 researchers studying social media data mentioned needing to work at multiple levels of aggregation, and having flexible units-of-analysis. Designers of coding software must be aware of their users' needs in this area, or be flexible to multiple options.

Depending on the meaning of the codes and the structure of the code scheme, it may be desirable for codes to be applied in a mutually-exclusive fashion, or to overlap and coincide. That is, coders might be able to apply only one code per message, or multiple codes per message. Further complicating matters, in some schemes there may be groups of mutually-exclusive codes along

with other non-mutually-exclusive codes. It is unclear what level of enforcement coding software should provide for such conventions. In our project, we interpreted affect as a dynamic, multi-faceted phenomenon, so coders could apply multiple affect codes per message. However, this flexibility came at the cost of complicating later analysis. For example, as I explained in more detail in Chapter 4, most methods of assessing inter-rater agreement assume that codes are applied in a mutually-exclusive fashion; to measure coding reliability in our project, we had to develop a custom version of *kappa* that worked with our style of coding.

## 6.6 Conclusion

The research group discussed in this chapter produced a grounded taxonomy of affect in chat communication (Scott et al., 2012), a machine learning tool for classifying affect in chat (Chapter 4), and a tool for qualitatively coding chat messages. However, our goals also included social science questions about affect in online communication: how do different categories of affect relate to one another, and to events and collaborative dynamics in the chat log? As I discussed in Chapter 3, answering complex social science questions, such as these, requires considering the data from many angles. Qualitative analysis is useful for sorting and organizing the data into meaningful categories, and for generating hypotheses and local insight into the phenomenon of interest. However, these qualitative categories must be integrated into an overall approach that also uses visual analysis, statistics, and other quantitative techniques. In this chapter, I focused on tools for creating categories over large amounts of data, and in Chapter 4, we experimented with using machine learning to project those categories up to larger amounts of data. However, better tools are needed to help researchers visually and statistically understand these categories in relation to the entire dataset. Connecting qualitative categorical analysis to large-scale quantitative analysis and visualization is an open problem, both for methodologists and technology builders.

In this chapter, I provided a detailed account of how our research group approached the study of emotion and affect in a large, multi-year chat dataset. We applied a combination of qualitative grounded theory methods and machine learning analysis; to support our qualitative data analysis processes, we designed and built an experimental coding and chat visualization tool, which became Text Prizm, an efficient web-based coding application. We used Text Prizm to collaboratively code thousands of chat messages, creating a grounded framework for affect, as well as training data for machine learning algorithms. Based on this case study, I presented several implications for design of qualitative data analysis software to support collaborative coding of large online communication and social media datasets, and highlighted many important areas for future work.

# Chapter 7: Conclusions

Datasets collected from online communication and social media platforms provide new ways to study and understand human communication and behavior. Broad participation in this work from researchers across many disciplines is important for ensuring a rich and productive research community around online social data. The available technology for working with social media and online communication data creates barriers and friction, affecting how researchers can get access to data, conduct exploratory analysis, and effectively integrate qualitative and quantitative approaches. In order to create better tools and technologies, we need a better understanding of the needs and constraints of researchers working in this area. Below, I review key findings from the studies described in the previous chapters. I then revisit my research questions, and discuss high-level themes and implications of this work.

## 7.1 Key Findings

In this dissertation, I have taken a human-centered design approach to understand the design of new technology to support social science analysis of social media and online communication data. I have described a naturalistic study of data analysis practices, and three technology design and development projects.

In Chapter 3 of this dissertation, I presented an ethnographic study on the data analysis practices of social scientists working with social media data. Based on interviews with social scientists, I described processes of data collection and analysis, and discussed challenges and opportunities for technology design. The study revealed barriers to accessing and collecting high quality social media data, challenging characteristics of social media datasets, and essential analytical techniques

(e.g. search, reading, visualization, computational analytics, and qualitative coding). Once researchers collect social media data, there are many opportunities to improve the ecosystem of data analysis tools that social scientists rely on:

- Context is important for analyzing social media data. Tools could automatically gather and present contextual information, e.g. details about user accounts, hashtags, and events.
- Social media data has many facets. Tools should support integrated analysis for more aspects of the data, including text-based messages, quantitative metadata, images, links, and network-relationship data.
- Social media datasets are volatile over time and geography. Tools must allow analysis at multiple scales or levels, through zoom and flexible aggregation.
- Researchers frequently move between different data analysis tools. To minimize the friction of transformation and restructuring data, software should adhere to standard data formats and be robust to unexpected inputs.
- Mixed methods are powerful for studying social media data. Researchers need tools that support both qualitative and quantitative analysis of social media data.
- Researchers shift between open-ended exploration and structured confirmatory analysis. Designing software to better support these transitions is an open problem.

More research is needed to advance our understanding of the data analysis practices of researchers working with social media data, as well as the practices of analysts in general. Data analysis is a dynamic, unpredictable, iterative process, where the structure of the activity can shift rapidly from one moment to the next. In the context of long-term research projects, the evolution of data analysis practices remains poorly understood.

In Chapter 4, I discussed a case study applying machine learning technology, in combination with grounded theory methods, to analyze emotion in online communication. My collaborators and I carried out an open-ended qualitative analysis on a small portion of a chat message dataset, and then applied machine learning to project our qualitative analysis onto the full dataset. In an iterative process informed by our close analysis of the chat data, we developed a machine learning tool, *ALOE*, which achieved 70-80% accuracy on the most frequent types of affect. *ALOE*[9] is available as an open-sourced software project to support future research on affect and online communication.

Based on this work, I explored possible misalignments between classic supervised machine learning frameworks and qualitative approaches. Positivist concepts that are basic to machine learning, such as ground truth, accuracy, and independence, seem to be in contradiction with the grounded, interpretive methodology we used, but this case study showed how aspects of computational algorithms, such as transparency and learning from the researchers' corpus, could make automated approaches more appropriate for social scientists using qualitative and mixed methodologies. Participants in the ethnographic study in Chapter 3 use a variety of tools to study their data from different angles; for these researchers, trainable machine learning tools like *ALOE* that are easy to use and easy to understand could be impactful as an option for classifying data, somewhere in between manual interpretation of the text, and simple, but noisy, keyword searches.

Future work should continue to interrogate the potential role for machine learning and other automation technologies in "big data" social science research. Cross-disciplinary community building is an important aspect of this problem. Technological solutions can play a role by

[9] http://depts.washington.edu/hdsl/tools

facilitating transparency and understanding of computational techniques (e.g. through informative visualizations and interactive exploration tools), but only through collaboration between machine learning experts and social scientists can real progress be made in breaking down barriers to the development of more applicable and suitable machine learning technology.

Chapter 5 described the design and evaluation of Agave, a collaborative visual analytics system for exploring events in large Twitter datasets. Informed by the challenges our research group experienced working with Twitter data, Agave supports exploration through a combination of time series visualizations, searching and filtering, and comparison between subsets of the data. It also allows collaborative analysis through a system of annotations and threaded discussions. We conducted a qualitative evaluation of Agave with several social media researchers, and found the tool was able to help users effectively explore a large event-based Twitter dataset. The timeline visualization led users to focus on significant events, while the annotations and discussion posts created by other users often suggested novel lines of analysis that helped unfamiliar users get started. The dual filter interface supporting compare and contrast made it easy for researchers to focus and ask more specific questions. In our evaluation, users mistrusted the sentiment analysis results displayed in Agave, an important challenge for future work.

As we saw in Chapter 3, while social scientists use a variety of tools and techniques for visualization and exploration of their social media data, few of these tools are actually designed for this kind of data. Social media's idiosyncrasies and complexities challenge general purpose tools, impeding effective exploratory analysis. For example, tools like Tableau and Gephi do not give adequate access to the raw messages, which prevents effective meaning-making. By focusing specifically on social media in Agave, we were able to provide data-type-specific optimizations, such as the promotion of timeline visualizations, as well as filtering and ranking by appropriate

metadata. Further, we allow users to search raw social media messages, an important consideration for our users, who search and read tweets extensively in their sensemaking process.

Researchers working with social media data need more exploratory visual analytics tools that support social media metadata fields, while staying close to the text. The literature contains examples of visual analytics systems designed for social media data, which offer some of these features: the *Vox Civitas* (Diakopoulos et al., 2010) and twitInfo (Marcus et al., 2011) systems help journalists mine social media during events; the "Visual Backchannel" system supports real-time monitoring of Twitter topics over time (Dork et al., 2010); Mazumdar et al. developed a system for non-technical emergency responders (Mazumdar et al., 2012); Hubmann-Haidvogel et al. describe a dynamic topography technique that shows how social media topics change over time (Hubmann-Haidvogel et al., 2012). However, these systems have not been evaluated with social scientists. To carry these promising techniques through into practice, future work should longitudinally investigate how visual analytics tools for social media data can be used in real social science projects over a long period of time.

In Chapter 6, I provided a detailed account of the design process used by our research group, studying affect in the Supernova Factory chat log dataset, as we iteratively created and deployed software for collaborative qualitative coding. Providing an efficient way for researchers to collaborate in open-ended qualitative analysis, this tool enabled our group to build a grounded taxonomy of affect in our dataset, and to prepare a ground truth coded dataset of tens of thousands of chat messages in preparation for the machine learning experiments discussed in Chapter 4. Based on the challenges we encountered and our design decisions, I discussed the following implications for qualitative data analysis (QDA) tools:

- Bias in the selection of codes is a threat to coding validity. Designers should consider how QDA user interfaces affect this bias, and further research on this topic is needed.
- In large, collaborative coding projects, quality control is a major concern; tools must provide agreement or reliability statistics and visualizations to help researchers monitor and control coding quality.
- For online communication datasets, a significant volume of coding may be needed. QDA tools should include interaction techniques, e.g. keyboard-based interaction, that maximize coder efficiency.
- Context is key to interpreting online communication and social media data. QDA software may be able to surface contextual information relevant to the coding process, saving users from the interruption of looking up external details manually.
- Teams carrying out qualitative coding on a large dataset must coordinate their work, and tools must provide support for maintaining mutual awareness of collaborators.
- In qualitative research, changing and iterating on the coding scheme may require various coding restructuring operations (e.g. delete, merge, split codes). QDA software should support gradual development of code schemes, and help users understand restructurings.
- Qualitative researchers use a variety of structures and styles of coding, and often the type of coding varies over the course of a project. Coding software must support different coding approaches.
- Researchers studying online communication datasets have unique analytical goals and needs, and QDA tools must enable researchers to do custom analysis with coded data by providing data export options or transparent access to back-end data storage.

- Manual qualitative coding is done on small amounts of data, so for larger online communication datasets, some method for sampling is needed. QDA tools should help researchers develop and execute appropriate sampling strategies.

More empirical work is needed evaluating the impact of qualitative data analysis software on research processes and products, and how to design QDA tools that promote high quality qualitative research.

## 7.2 Challenges for Future Work

The goal of this dissertation has been to address the following three research questions:

- *What are the goals, barriers, tools, and processes associated with mixed methods online social data research?*
- *How can machine learning techniques be applicable for automation in mixed methods research with online communication data?*
- *How can tools be designed to help researchers explore and code large online communication and social media datasets?*

I have discussed the challenges and data analysis practices of social scientists working with social media data, explored design implications for machine learning, visualization, and qualitative coding tools, and developed a collection of open-sourced software tools for analyzing online communication and social media datasets. Below, I discuss broad themes related to data exploration and qualitative methods that emerged in this dissertation.

### 7.2.1 Data Analysis in Long-Term Projects

Research projects evolve over their lifetimes; while researchers begin with one set of questions and goals, over time the focus may change. Accompanying these changes, the interactions between researchers and their data change as well, involving a mix of data collection, cleaning, exploration, analysis, and presentation activities. Yet, most of the tools commonly used to work with data do little to ease these "phase transitions" in research; instead, many data analysis tools support only narrow, short-term transactions with the data. For example, if a Tableau workbook was created while exploring a dataset, what role does it play, months later, when a new set of charts are needed to prepare a paper? How does a spreadsheet created to read and code a group of random tweets early in a project inform the eventual construction of new queries for a more focused analysis? The work of tying together data, information, and knowledge resources across the lifetime of a project as the research questions and findings converge is a challenge that many current data analysis tools fail to address, instead leaving such work to the researcher.

How can tools help make this long-term process more efficient? The interwoven texture of data interactions in research projects is complex and difficult to characterize, but one thread that has repeatedly surfaced in this dissertation is data *exploration*. For the social scientists that I have worked with, exploration and exploratory data analysis was an integral and instrumental to idea development and convergence; through exploration, researchers gradually refined their focus, and repeatedly proposed, developed, and discarded hypotheses and theoretical constructs. While researchers approach projects with preconceived ideas in mind, exploratory data analysis can shift attention to unexpected places and suggest new hypotheses or theories. Because of the complexity of social media and online datasets, even confirmatory analysis, where researchers are testing or

evaluating an idea against their data, takes on the guise of exploration, since researchers consider their data from many angles using a variety of tools and techniques.

As I have argued in this dissertation, exploration of social media and online communication datasets is both critical and time consuming. The interviewees in Chapter 3 stressed the importance of exploration, but explained that they had to spend a great deal of time gathering extensive outside information even to understand what individual social media messages mean. Similarly, in Chapters 4 and 6, our work depended on understanding the context and background of the dataset, primarily through close reading and coding. In this domain, qualitative analysis can support effective exploration. Qualitative methods are sometimes thought of as inherently exploratory, and in mixed methods projects, they tend to fill an exploratory role (Creswell, 2014). In the grounded theory approach mentioned by some participants in Chapter 3 and used with the Supernova Factory chat dataset in Chapters 4 and 6, exploratory qualitative analysis was used to iteratively and inductively construct theoretical concepts of affect and emotion in online chat communication.

Better tools to enable exploration of this type of data are needed. Specific challenges with the process of data exploration and gradual development and convergence of research findings arose in several contexts. First, when researchers are already struggling with a mix of technologies to formulate and examine new views of their datasets during exploration, it is challenging to simultaneously generate, record, and track ideas and findings. In Chapter 5, I discussed how the threaded discussions built into Agave could help users record and share findings with collaborators, facilitating the accumulation of evidence around discussion themes and topics of interest to the group. Related work on collaborative sensemaking and visual analytics contains other ideas for how to support knowledge construction processes in visualization and data analysis

tools (Fischer, Bruhn, Gräsel, & Mandl, 2002; Heer & Agrawala, 2008; Klein et al., 2006; Pirolli & Card, 2005).

Second, the process of restructuring qualitative coding schemes over time can be complex and challenging. In Chapter 3, some participants described using open-ended coding early on in their projects; while this early coding work might be discarded, the initial attempts can eventually inform systematic coding processes later on. In Chapter 6, I discussed how our research group also struggled with managing these transitions as we iteratively reorganized our coding scheme, converging from a chaotic initial set of over 60 categories to about 30 affect codes, and proposed ways in which collaborative qualitative coding tools could make changing and refining codes easier. While traditional QDA software does typically provide support for merging, renaming, restructuring, and refining codes (Bringer et al., 2006; St John & Johnson, 2000), some qualitative researchers claim the definitiveness and finality of these actions contradicts the subtle development of understanding they find desirable for their work (Goble et al., 2012). Moreover, the spreadsheets that many researchers use for coding large social media datasets do not provide even these features.

Future work should approach research projects as dynamic, collaborative activities that take place over a long time span, and explore related design problems such as managing the development of qualitative coding schemes and facilitating exploratory analysis of social media and online communication datasets.

### 7.2.2 Qualitative and Mixed Methods for Big Data

The term "big data" often presupposes a quantitative approach to data analysis, and in many scientific domains, that assumption is reasonable. On very large datasets, computational processing is required, and computers work with numbers; for summarizing and finding patterns,

quantitative statistics and visualizations are highly effective. However, some researchers have been critical of the focus on quantity and quantification in the big data movement (boyd & Crawford, 2012), and in the social sciences (Latour, 2010). Many questions of great significance cannot easily be answered quantitatively. If quantitative big data research becomes more dominant in the sciences, how does that limit what we can learn and how we can learn it? Qualitative methods, while labor intensive and focused on smaller amounts of data, are powerful in ways that quantitative methods are weak (Creswell, 2014), but what role can qualitative approaches play for studying big data?

There are many challenges to be considered. Qualitative analysis relies on expensive manual labor to read and interpret rich data; it is not usually cost-effective to apply such a process to large online social datasets (Rosé et al., 2008). If only a small amount of data can be analyzed closely, selecting which data to focus on requires care; techniques such as clustering and visualization may be useful here, as they can show some types of patterns, groups, and other structures in the data, prior to manual analysis. Once data is selected, it is often analyzed using coding, but this can be a time-consuming, tedious process. Interviewees in Chapter 3 struggled to apply qualitative coding because of tools that did not work well with social media data. In Chapter 6, I explained how our research group decided it would be more efficient to design and build our own specialized coding tool to analyze a large amount of chat log data. Our coding tool, Text Prizm, presents chat messages in an easy-to-interpret format, accelerates coding through easy keyboard interactions, and supports collaboration, enabling users to reach more data. However, many opportunities exist to improve upon these processes, e.g. integrating contextual information about messages, helping researchers debug and visualize their coding, and providing tools to refine the coding scheme.

Better tools for qualitative coding of short online messages are needed, but coding is only part of the picture: in most big online social data projects, qualitative analysis must also play well with quantitative and computational analysis. In this dissertation, many of the researchers I worked with preferred to use both qualitative and quantitative techniques, as part of a mixed methods approach. The strategies used in mixed methods research vary: qualitative exploration can be followed by broad quantitative confirmation, quantitative testing can precede explanatory qualitative case studies, or more concurrent and integrated strategies can be used (Creswell, 2014). Mixed methods are thought to combine many of the benefits of both quantitative and qualitative research (Johnson & Onwuegbuzie, 2004), e.g. the precision of numbers, with the illustrative power and depth of words, pictures, and narratives. Perhaps it is the complexity of these datasets, combining unstructured text and a dizzying array of metadata, which makes mixed methods appealing.

Some of the challenges here are methodological. Some qualitative practitioners and theorists have argued that the underlying thought systems of qualitative research (e.g. constructivism or interpretivism) are incompatible with positivist and post-positivist quantitative paradigms (Creswell, 2014; Johnson & Onwuegbuzie, 2004). The perceived "incompatibility" between qualitative and quantitative approaches has a long history, and has even been extended to the point that some qualitative researchers are suspicious of using any type of research software (Banner & Albarrran, 2009; Coffey et al., 1996; Goble et al., 2012). Advocates of mixed methods focus on the strengths, weaknesses, and commonalities of different methodologies, with pragmatism as the underlying thought system (Johnson & Onwuegbuzie, 2004).

With these ongoing methodological debates as background, social scientists who are mixing methods to study online social data must find feasible ways to resolve the diverse languages, practices, and products of their approaches. One strategy I have discussed here is to use machine

learning techniques, such as clustering (Janasik et al., 2009), or supervised classification (Rosé et al., 2008). Some of my interviewees in Chapter 3 discussed using clustering as a precursor to qualitative analysis; for these researchers, unsupervised computational analysis produced ideas and concepts that could guide selection and focused coding of smaller amounts of data. In Chapters 4 and 6, I discussed integrating qualitative and large-scale quantitative analysis through supervised classification. We used Text Prizm to build a grounded framework for affective communication in a chat dataset, and collaboratively coded thousands of lines of chat. Exporting our coded data, we trained classifiers with the goal of automating coding on the full dataset.

While better machine learning tools and other computational analytics could be useful in qualitative and mixed methods research with online social data, it is important not to lose sight of the importance and value of human expertise. The questions that social scientists ask are complex and difficult to reduce to computational abstractions. Thus, I would argue that the most impactful solutions in this area will be human-driven, striking a balance between the strengths of expert researchers (e.g. domain knowledge and human intuition) and the efficiency of computational data processing. Good interaction and data visualization design are the key to making this vision a reality. How could interactive machine learning, e.g. (Amershi et al., 2012; Ware et al., 2001), be used as an integral part of qualitative analysis? As with Agave in Chapter 5, how can we develop better methods of visualizing unstructured text datasets to help researchers find and focus on more important segments of data, e.g. (Bhowmick, 2006; Don et al., 2007; Fang, Lwin, & Ebright, 2006; Risch, Kao, Poteet, & Wu, 2008)? Several examples of social media and online communication research that pursue other approaches for integrating qualitative and quantitative methods were reviewed in Chapter 2, but there are many open questions to explore.

## 7.3 Strategies and Recommendations

Over the course of this dissertation, I have had the opportunity to work with many excellent researchers, and witnessed many struggles and successes. I will close with some thoughts for practitioners of online social data research, and for technology developers and designers.

Acquiring programming and data science skills is a major barrier for researchers working in this interdisciplinary space. For those who wish to study online social data but do not know programming, databases, or statistics (collectively, data science skills), I argue that it is worth the effort to learn. Even when researchers are collaborating with others who have data science skills, those who do not have the ability to directly analyze their data may experience a frustrating sense of distance, and may not be able to work to their full potential. For example, being dependent on other collaborators to retrieve information from a database can be a major source of friction; acquiring competency with a few types of database queries can be tremendously empowering.

However, learning these technologies is not easy. The researchers I worked with amazed me with their resourcefulness and their drive to find and learn new tools. I have observed and heard about countless skill shares and workshops, where researchers help each other learn various data science techniques. If learning these skills individually is too daunting, as it is for many people, beginning researchers may reach out to find collaborators and community resources. The popularity of studying online social data makes it easier than ever to find others with shared learning interests; several of my participants were part of informal learning communities and groups which supported their efforts to acquire programming and data science skills. There are hundreds and hundreds of tools and technologies available which may be useful for analyzing online social data; for many,

the most effective way to find and learn these tools will be through community and collaboration. Faculty should encourage students working with online social data to seek out these resources.

Within research projects, collaboration is also a powerful strategy for success. For many students, it is infeasible to develop deep competency in multiple disciplines, e.g. both computer science and political science. But, it may be possible more many researchers to acquire enough working knowledge of other disciplines to support interdisciplinary collaboration, and then to team up with experts from several disciplines to ask interesting questions about online social datasets. There are challenges to interdisciplinary collaboration in data-intensive science, including language barriers (Monteiro & Keating, 2009), and data disagreements (Edwards et al., 2011), but many of my colleagues and participants were part of successful research teams composed of students from multiple disciplines.

For those designing and building the next generation of tools for online social data research, I stress the importance of working closely with users and listening to their needs. It is all too easy for technology builders to fall back on assumptions about users' needs and pain points, and to become fascinated by technical ingenuity. For tool builders who are technology researchers, technology research goals, such as novelty, can sometimes be a distraction from solving users' most important problems. Bloom has pointed out that pressure to publish in researchers' "home" disciplines can create conflicts in interdisciplinary collaborative projects (Bloom, 2013). On the other hand, others argue that applied research and design projects make important research contributions (Zimmerman et al., 2007), so perhaps the apparent conflict between good design and good research is an illusion. In any case, I have found that it is helpful to have an upfront, open discussion about the research priorities of tool development projects, and to keep these priorities explicit in decision-making. For those creating new tools as part of research projects, and who

wish to take a human-centered approach, applying participatory and observational methods very early in the ideation process can deliver great value. Frequent informal testing of prototypes with sympathetic users throughout the design and implementation of solutions can also provide crucial course corrections and ensure that users' voices are not lost. And, eventually, to ensure that tools are polished enough to make the leap into practice, they should be provided to users for further testing and refinement in their daily work.

Building tools where the users are researchers also introduces special challenges. The social scientists that I worked with in this dissertation were doing very different research, and a review of the published literature in this area will show tremendous diversity in interests, methods, and data sources. Some have argued that software in data-intensive scientific research must take a holistic, integrated approach to support data management and analysis end-to-end (Howe et al., 2008). However, in the diverse social science domain discussed here, my colleagues and I often struggled to identify cross-cutting problems which we could solve with reusable tools. While opportunities for complex, end-to-end tools may exist, I recommend against big, many-featured tools that attempt to solve too many problems. Such tools are difficult to prototype, difficult to explain to users, expensive to build, and hard to evaluate; most likely, they will either meet all needs for only a tiny number of users, or meet only a few needs for a larger number of users. Researchers working with online social data are already using a large number of tools, and replacing this ecosystem with a single complex tool is an ambitious endeavor for researchers with limited resources. I believe that a more promising approach is to create smaller, streamlined tools that focus on one problem, and which play well with the ecosystem of other commonly-used tools.

With data-intensive approaches taking hold across the sciences more than ever before, software has become a critical component of the research process. Yet, "researchers should pause and

consider that technology is more than a tool [... it] requires researchers to reframe ideas about what can be done and how it is done [and] may have predetermined what is drawn to the researcher's attention" (St John & Johnson, 2000). From this point of view, the design and selection of software becomes as intellectually significant to the researcher as the choice of method. My goal in this dissertation has been to explore the design space for software that social scientists could use to analyze large online communication and social media datasets. With big data it is tempting to focus only on the numbers, but online social data is deeply human. Therefore, it is my hope that this dissertation encourages the creation of more and better tools enabling qualitative and mixed methods approaches for analyzing online social data, and that more social scientists will consider adopting tools to explore their data from these perspectives.

# References


Agarwal, S. D., Bennett, W. L., Johnson, C. N., & Walker, S. (2014). A Model of Crowd Enabled Organization: Theory and Methods for Understanding the Role of Twitter in the Occupy Protests. *International Journal of Communication*, *8*. Retrieved from http://ijoc.org/index.php/ijoc/article/view/2068

Amabile, T. M., Barsade, S. G., Mueller, J. S., & Staw, B. M. (2005). Affect and Creativity at Work. *Administrative Science Quarterly*, *50*(3), 367–403. http://doi.org/10.2189/asqu.2005.50.3.367

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *Text, Speech and Dialogue*. Pilsen, Czech Republic: Springer.

Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., & Suh, J. (2015). ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 337–346). Seoul, Korea: ACM. http://doi.org/10.1145/2702123.2702509

Amershi, S., Fogarty, J., Kapoor, A., & Tan, D. (2011). Effective End-User Interaction with Machine Learning. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*.

Amershi, S., Fogarty, J., & Weld, D. (2012). Regroup: Interactive Machine Learning for On-demand Group Creation in Social Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 21–30). Austin, TX, USA: ACM. http://doi.org/10.1145/2207676.2207680

Aragon, C. R., Poon, S. S., Monroy-Hernández, A., & Aragon, D. (2009). A Tale of Two Online Communities: Fostering Collaboration and Creativity in Scientists and Children. In *Proceedings of the 7th ACM Conference on Creativity and cognition (C&C)* (pp. 9–18). Berkeley, CA: ACM. http://doi.org/10.1145/1640233.1640239

Ashforth, B. E., & Humphrey, R. H. (1995). Emotion in the Workplace: A Reappraisal. *Human Relations*, *48*(2), 97–125. http://doi.org/10.1177/001872679504800201

Banner, D. J., & Albarrran, J. W. (2009). Computer-assisted qualitative data analysis software: a review. *Canadian Journal of Cardiovascular Nursing*, *19*(3), 24–31.

Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, *47*(4), 644–675.

Bennett, W. L., Segerberg, A., & Walker, S. (2014). Organization in the crowd: peer production in large-scale networked protests. *Information, Communication & Society*, *17*(2), 232–260. http://doi.org/10.1080/1369118X.2013.870379

Bergman, M. M. (2008). The straw men of the qualitative-quantitative divide and their influence on mixed methods research. In *Advances in Mixed Methods Research* (pp. 11–21). SAGE Publications, Inc.

Bhowmick, T. (2006). Building an Exploratory Visual Analysis Tool for Qualitative Researchers. In *AutoCarto*. Vancouver, WA.

Bjork, S., & Redström, J. (2000). Redefining the focus and context of focus+context visualization. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*. Salt Lake City, UT, USA: IEEE. http://doi.org/10.1109/INFVIS.2000.885094

Blomberg, J., Giacomi, J., Mosher, A., & Swenton-Wall, P. (1993). Ethnographic field methods and their relation to design. In D. Schuler & A. Namioka (Eds.), *Participatory design: Principles and practices* (pp. 123–155). Lawrence Erlbaum Associates.

Bloom, J. (2013, November 25). Novelty Squared: A Challenge of Modern Interdisciplinary Scientific Collaboration. Retrieved September 16, 2015, from https://medium.com/@profjsb/novelty-squared-dd88857f662

Blum, A. (1998). On-line algorithms in machine learning. In A. Fiat & G. J. Woeginger (Eds.), *Online Algorithms*. Springer Berlin Heidelberg.

Boehner, K., Depaula, R., Dourish, P., & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies*, *65*(4), 275–291. http://doi.org/10.1016/j.ijhcs.2006.11.016

Bogdan, R. C., & Biklen, S. K. (1997). *Qualitative Research for Education: An Introduction to Theory and Methods* (Third Edition). Boston: Allyn & Bacon.

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. http://doi.org/10.1016/j.jocs.2010.12.007

boyd, danah. (2007). Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life. In D. Buckingham (Ed.), *Youth, Identity, and Digital Media*. Cambridge, MA: The MIT Press.

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. http://doi.org/10.1080/1369118X.2012.678878

boyd, danah, & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. http://doi.org/10.1111/j.1083-6101.2007.00393.x

boyd, danah, Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *Proceedings of the 43rd Hawaii International Conference on System Sciences (HICSS)*. Honolulu, HI: IEEE. http://doi.org/10.1109/HICSS.2010.412

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. http://doi.org/10.1016/S0031-3203(96)00142-2

Brehmer, M., Carpendale, S., Lee, B., & Tory, M. (2014). Pre-design Empiricism for Information Visualization: Scenarios, Methods, and Challenges. In *Proceedings of the 5th Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV)* (pp. 147–151). Paris, France: ACM. http://doi.org/10.1145/2669557.2669564

Bringer, J. D., Johnston, L. H., & Brackenridge, C. H. (2006). Using Computer-Assisted Qualitative Data Analysis Software to Develop a Grounded Theory Project. *Field Methods*, *18*(3), 245–266. http://doi.org/10.1177/1525822X06287602

Brooks, M., Amershi, S., Lee, B., Drucker, S. M., Kapoor, A., & Simard, P. (2015). FeatureInsight: Visual Support for Error-Driven Feature Ideation in Text Classification. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*. IEEE.

Brooks, M., Kuksenok, K., Torkildson, M. K., Perry, D., Robinson, J. J., Scott, T. J., … Aragon, C. R. (2013). Statistical Affect Detection in Collaborative Chat. In *Proceedings of the 2013*

*Conference on Computer Supported Cooperative Work (CSCW)* (pp. 317–328). San Antonio, TX, USA: ACM. http://doi.org/10.1145/2441776.2441813

Brooks, M., Robinson, J. J., Torkildson, M. K., Hong, S. (Ray), & Aragon, C. R. (2014). Collaborative Visual Analysis of Sentiment in Twitter Events. In *The 11th International Conference on Cooperative Design, Visualization, and Engineering*. Seattle, WA: Springer.

Buchanan, R. (1992). Wicked Problems in Design Thinking. *Design Issues*, *8*(2), 5–21. http://doi.org/10.2307/1511637

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE Publications, Inc.

Charmaz, K., & Belgrave, L. L. (2002). Qualitative interviewing and grounded theory analysis. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE Handbook of Interview Research: The Complexity of the Craft*. SAGE Publications, Inc.

Chin, G., Jr., Kuchar, O. A., & Wolf, K. E. (2009). Exploring the Analytical Processes of Intelligence Analysts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 11–20). New York, NY, USA: ACM. http://doi.org/10.1145/1518701.1518704

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 74–77). ACM.

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 443–452). ACM.

Coffey, A., Holbrook, B., & Atkinson, P. (1996). Qualitative Data Analysis: Technologies and Representations. *Sociological Research Online*, *1*(1). Retrieved from http://www.socresonline.org.uk/1/1/4.html

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*. http://doi.org/10.1177/001316446002000104

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications, Inc.

Crowston, K., Liu, X., & Allen, E. E. (2010). Machine learning and rule-based automated coding of qualitative data. *Proceedings of the American Society for Information Science and Technology Annual Meeting*, *47*(1), 1–2. http://doi.org/10.1002/meet.14504701328

Crowston, K., & Nahon, K. (2015). Introduction to the Digital and Social Media Track. In *Proceedings of the 48th Hawaii International Conference on System Sciences (HICSS)*. Honolulu, HI, USA: IEEE. http://doi.org/10.1109/HICSS.2015.187

Dailey, D., & Starbird, K. (2014). Visible Skepticism: Community Vetting after Hurricane Irene. In *Proceedings of the 11th International ISCRAM Conference*. Retrieved from http://iscram2014.ist.psu.edu/sites/default/files/misc/proceedings/p178.pdf

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Major Life Changes and Behavioral Markers in Social Media: Case of Childbirth. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW)* (pp. 1431–1442). San Antonio, TX, USA: ACM. http://doi.org/10.1145/2441776.2441937

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*. The AAAI Press. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6124

Denzin, N. K., & Lincoln, Y. S. (2011). The discipline and practice of qualitative research. In *The SAGE Handbook of Qualitative Research* (Fourth Edition). Thousand Oaks, CA, USA: SAGE Publications, Inc.

Diakopoulos, N., De Choudhury, M., & Naaman, M. (2012). Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2451–2460). ACM.

Diakopoulos, N., Naaman, M., & Kivran-Swaine, F. (2010). Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)* (pp. 115–122). Salt Lake City, UT, USA: IEEE. http://doi.org/10.1109/VAST.2010.5652922

Diesner, J., Frantz, T. L., & Carley, K. M. (2005). Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different." *Computational & Mathematical Organization Theory*, *11*(3), 201–228. http://doi.org/10.1007/s10588-005-5377-0

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE*, *6*(12), e26752. http://doi.org/10.1371/journal.pone.0026752

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. http://doi.org/10.1145/2347736.2347755

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., … Plaisant, C. (2007). Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)* (pp. 213–222). Lisbon, Portugal: ACM. http://doi.org/10.1145/1321440.1321473

Dork, M., Gruen, D., Williamson, C., & Carpendale, S. (2010). A Visual Backchannel for Large-Scale Events. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1129–1138. http://doi.org/10.1109/TVCG.2010.129

Dourish, P., & Bellotti, V. (1992). Awareness and Coordination in Shared Workspaces. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work* (pp. 107–114). Toronto, Ontario: New York: ACM.

Dourish, P., & Chalmers, M. (1994). Running out of space: Models of information navigation. In *Proceedings of HCI '94*. Glasgow, Scotland: ACM Press. Retrieved from http://www.dourish.com/publications/1994/hci94-navigation.pdf

Driscoll, K., & Walker, S. (2014). Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication*, *8*(0), 20.

Drisko, J. (2006). QDA Software. Retrieved August 8, 2015, from http://sophia.smith.edu/~jdrisko/qdasoftw.htm

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science*, *41*(5), 667–690. http://doi.org/10.1177/0306312711413314

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing Ethnographic Fieldnotes* (Second Edition). Chicago: University Of Chicago Press.

Fang, S., Lwin, M., & Ebright, P. (2006). Visualization of unstructured text sequences of nursing narratives. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC)* (pp. 240–244). Dijon, France: ACM. http://doi.org/10.1145/1141277.1141331

FAQs about adding location to your Tweets. (2015). Retrieved September 19, 2015, from https://support.twitter.com/articles/78525

Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction*, *12*(2), 213–232. http://doi.org/10.1016/S0959-4752(01)00005-6

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. http://doi.org/10.1037/h0031619

Frayling, C. (1994). Research in Art and Design. *Royal College of Art Research Papers*, *1*(1). Retrieved from http://researchonline.rca.ac.uk/id/eprint/384

Geertz, C. (1973). *The Interpretation Of Cultures*. New York: Basic Books.

Gilbert, E. (2012). Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)* (pp. 1037–1046). Seattle, WA: ACM Press. http://doi.org/10.1145/2145204.2145359

Gill, A. J., French, R. M., Gergle, D., & Oberlander, J. (2008). The language of emotion in short blog texts. In *Proceedings of the ACM 2008 conference on Computer Supported Cooperative Work (CSCW)*. http://doi.org/10.1145/1460563.1460612

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction.

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. Stanford University. Retrieved from http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

Goble, E., Austin, W., Larsen, D., Kreitzer, L., & Brintnell, E. S. (2012). Habits of Mind and the Split-Mind Effect: When Computer-Assisted Qualitative Data Analysis Software is Used in Phenomenological Research. *Forum: Qualitative Social Research*, *13*(2). Retrieved from http://nbn-resolving.de/urn:nbn:de:0114-fqs120227

Goggins, S. P., Laffey, J., & Gallagher, M. (2011). Completely online group formation and development: small groups as socio-technical systems. *Information Technology & People*, *24*(2), 104–133. http://doi.org/10.1108/09593841111137322

Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516–539. http://doi.org/10.1002/asi.22802

Golder, S. A., & Macy, M. W. (2011). Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, *333*(6051), 1878–1881. http://doi.org/10.1126/science.1202775

Grandey, A. (2008). Emotions at Work: A Review and Research Agenda. In J. Barling & C. L. Cooper (Eds.), *Handbook of Organizational Behavior*. London: SAGE Publications, Inc.

Guzman, J., & Poblete, B. (2013). On-line Relevant Anomaly Detection in the Twitter Stream: An Efficient Bursty Keyword Detection Model. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description (ODD)*. Chicago, IL: ACM. http://doi.org/10.1145/2500853.2500860

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10. http://doi.org/10.1145/1656274.1656278

Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L., & Hsu, M.-C. (2011). Visual sentiment analysis on twitter data streams. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 277–278). Providence, RI: IEEE. http://doi.org/10.1109/VAST.2011.6102472

Havre, S., Hetzler, E., Whitney, P., & Nowell, L. (2002). Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 9–20. http://doi.org/10.1109/2945.981848

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, *1*(1), 77–89. http://doi.org/10.1080/19312450709336664

Heer, J., & Agrawala, M. (2008). Design Considerations for Collaborative Visual Analytics. *Information Visualization*, *7*(1), 49–62. http://doi.org/10.1057/palgrave.ivs.9500167

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. http://doi.org/10.1145/1753326.1753357

Heer, J., Viegas, F., & Wattenberg, M. (2007). Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. San Jose, CA: ACM. http://doi.org/10.1145/1240624.1240781

Heimerl, F., Jochim, C., Koch, S., & Ertl, T. (2012). FeatureForge: A Novel Tool for Visually Supported Feature Engineering and Corpus Revision. In *Proceedings of COLING 2012: Posters* (pp. 461–470). Mumbai, India: The COLING 2012 Organizing Committee.

Howe, B., Lawson, P., Bellinger, R., Anderson, E., Santos, E., Freire, J., … Silva, C. (2008). End-to-End eScience: Integrating Workflow, Query, Visualization, and Provenance at an Ocean Observatory. In *IEEE 4th International Conference on eScience* (pp. 127–134). Indianapolis, IN. http://doi.org/10.1109/eScience.2008.67

Huberman, B. A. (2012). Sociology of science: Big data deserve a bigger audience. *Nature*, *482*(7385), 308–308. http://doi.org/10.1038/482308d

Hubmann-Haidvogel, A., Brasoveanu, A. M., Scharl, A., Sabou, M., & Gindl, S. (2012). Visualizing contextual and dynamic features of micropost streams. In *Proceedings of the WWW 2012 Workshop on Making Sense of Microposts (MSM)*. Lyon, France. Retrieved from http://eprints.weblyzard.com/36/

Humble, Á. M. (2012). Qualitative Data Analysis Software: A Call for Understanding, Detail, Intentionality, and Thoughtfulness. *Journal of Family Theory & Review*, *4*(2), 122–137. http://doi.org/10.1111/j.1756-2589.2012.00125.x

Janasik, N., Honkela, T., & Bruun, H. (2009). Text Mining in Qualitative Research: Application of an Unsupervised Learning Method. *Organizational Research Methods*, *12*(3), 436–460. http://doi.org/10.1177/1094428108317202

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Proceedings of the 10th European Conference on Machine Learning (ECML)* (pp. 137–142). Chemnitz, Germany: Springer Berlin Heidelberg. http://doi.org/10.1007/BFb0026683

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, *33*(7), 14–26. http://doi.org/10.3102/0013189X033007014

Jones, M. (2007). Using software to analyse qualitative data. *Faculty of Commerce - Papers (Archive)*. Retrieved from http://ro.uow.edu.au/commpapers/429

Jones, R. H. (2004). The problem of context in computer-mediated communication. In P. LeVine & R. Scollon (Eds.), *Discourse and technology: Multimodal discourse analysis* (pp. 20–33). Georgetown University Press.

Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise Data Analysis and Visualization: An Interview Study. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2917–2926. http://doi.org/10.1109/TVCG.2012.219

Kang, Y., & Stasko, J. (2011). Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In *IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 21–30). Providence, RI: IEEE. http://doi.org/10.1109/VAST.2011.6102438

Keim, D. A., Kohlhammer, J., Ellis, G., & Mansmann, F. (Eds.). (2010). *Mastering The Information Age-Solving Problems with Visual Analytics*. Eurographics Association.

Keshtkar, F., & Inkpen, D. (2009). Using sentiment orientation features for mood classification in blogs. In *Proceedings of the International Conference on Natural Language Processing and*

*Knowledge Engineering (NLP-KE)*. Dalian: IEEE. http://doi.org/10.1109/NLPKE.2009.5313734

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 453–456). Florence, Italy: ACM Press. http://doi.org/10.1145/1357054.1357127

Kivran-Swaine, F., & Naaman, M. (2011). Network Properties and Social Sharing of Emotions in Social Awareness Streams. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW)* (pp. 379–382). Hangzhou, China: ACM. http://doi.org/10.1145/1958824.1958882

Klein, G., Moon, B., & Hoffman, R. R. (2006). Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems*, *21*(5), 88–92. http://doi.org/10.1109/MIS.2006.100

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. http://doi.org/10.1073/pnas.1320040111

Kulesza, T., Amershi, S., Caruana, R., Fisher, D., & Charles, D. (2014). Structured Labeling for Facilitating Concept Evolution in Machine Learning. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI)* (pp. 3075–3084). Toronto, Canada: ACM. http://doi.org/10.1145/2556288.2557238

Kulesza, T., Wong, W.-K., Stumpf, S., Perona, S., White, R., Burnett, M. M., … Ko, A. J. (2009). Fixing the Program My Computer Learned: Barriers for End Users, Challenges for the Machine. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI)*. Haifa, Israel: ACM. http://doi.org/10.1145/1502650.1502678

Lafferty, N. T., & Manca, A. (2015). Perspectives on social media in and as research: A synthetic review. *International Review of Psychiatry*, *27*(2), 85–96. http://doi.org/10.3109/09540261.2015.1009419

Lane, D. M., Napier, H. A., Peres, S. C., & Sandor, A. (2005). Hidden Costs of Graphical User Interfaces: Failure to Make the Transition from Menus and Icon Toolbars to Keyboard

Shortcuts. *International Journal of Human-Computer Interaction*, *18*(2), 133–144. http://doi.org/10.1207/s15327590ijhc1802_1

Latour, B. (2010). Tarde's idea of quantification. In M. Candea (Ed.), *The Social After Gabriel Tarde: Debates and Assessments* (pp. 145–162). London: Routledge.

Lethbridge, T. C., Sim, S. E., & Singer, J. (2005). Studying Software Engineers: Data Collection Techniques for Software Field Studies. *Empirical Software Engineering*, *10*(3), 311–341. http://doi.org/10.1007/s10664-005-1290-x

Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI)* (pp. 125–132). http://doi.org/10.1145/604045.604067

Luff, P., Hindmarsh, J., & Heath, C. (2000). *Workplace Studies: Recovering Work Practice and Informing System Design*. Cambridge University Press.

Maddock, J., Starbird, K., Al-Hassani, H. J., Sandoval, D. E., Orand, M., & Mason, R. M. (2015). Characterizing Online Rumoring Behavior Using Multi-Dimensional Signatures. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)* (pp. 228–241). Vancouver, BC: ACM. http://doi.org/10.1145/2675133.2675280

Manovich, L. (2011). Trending: the promises and the challenges of big social data. In M. K. Gold, *Debates in the Digital Humanities*. University of Minnesota Press.

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011). Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 227–236). ACM. http://doi.org/10.1145/1978942.1978975

Mark, G., Fuchs, L., & Sohlenkamp, M. (1997). Supporting groupware conventions through contextual awareness. In *Proceedings of The 5th European Conference on Computer Supported Cooperative Work (ECSCW)* (pp. 253–268). Springer.

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*. http://doi.org/10.3758/s13428-011-0124-6

Mazumdar, S., Ciravegna, F., Gentile, A. L., & Lanfranchi, V. (2012). Visualising context and hierarchy in social media. In *International EKAW Workshop on Intelligent Exploration of Semantic Data (IESD)* (Vol. 2012). Galway City, Ireland. Retrieved from http://imash.leeds.ac.uk/event/pdf/Mazumdar_7.pdf

McLafferty, E., & Farley, A. H. (2006). Analysing qualitative research data using computer software. *Nursing Times, 102*(24), 34–36.

Mentis, H. M., Reddy, M., & Rosson, M. B. (2010). Invisible emotion. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW)* (pp. 311–320). Savannah, GA: ACM Press. http://doi.org/10.1145/1718918.1718975

Metzger, N., & Zare, R. N. (1999). Interdisciplinary Research: From Belief to Reality. *Science, 283*(5402), 642–643. http://doi.org/10.1126/science.283.5402.642

Millen, D. R. (2000). Rapid Ethnography: Time Deepening Strategies for HCI Field Research. In *Proceedings of the 3rd Conference on Designing Interactive Systems (DIS)* (pp. 280–286). Brooklyn, NY: ACM. http://doi.org/10.1145/347642.347763

Milliken, F. J., Bartel, C. A., & Kurtzberg, T. R. (2003). Diversity and creativity in work groups: A dynamic perspective on the affective and cognitive processes that link diversity and performance. In P. B. Paulus & B. A. Nijstad (Eds.), *Group creativity: Innovation through collaboration* (pp. 32–62). New York: Oxford University Press.

Mishne, G. (2005). Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access (Style)*. Retrieved from http://hdl.handle.net/11245/1.253394

Mohammad, S. (2012). Portable Features for Classifying Emotional Text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)* (pp. 587–591). Montreal, Canada.

Monteiro, M., & Keating, E. (2009). Managing Misunderstandings: The Role of Language in Interdisciplinary Scientific Collaboration. *Science Communication*. http://doi.org/10.1177/1075547008330922

Moore, B. S., & Isen, A. M. (1990). *Affect and Social Behavior*. Cambridge University Press.

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of Social Media Research: Common Concerns and Practical Considerations. *Cyberpsychology, Behavior and Social Networking*, *16*(9), 708–713. http://doi.org/10.1089/cyber.2012.0334

Munzner, T. (2009). A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 921–928. http://doi.org/10.1109/TVCG.2009.111

Murdock Jr., B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488. http://doi.org/10.1037/h0045106

Murthy, D. (2008). Digital Ethnography An Examination of the Use of New Technologies for Social Research. *Sociology*, *42*(5), 837–855. http://doi.org/10.1177/0038038508094565

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2010). Affect Analysis Model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, *17*(01), 95–135. http://doi.org/10.1017/S1351324910000239

Norman, D. (2013). *The Design of Everyday Things* (Revised Edition). New York, New York: Basic Books.

Orlikowski, W. J. (2000). Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science*, *11*(4), 404–428. http://doi.org/10.1287/orsc.11.4.404.14600

Patel, K., Fogarty, J., Landay, J. A., & Harrison, B. (2008). Investigating Statistical Machine Learning As a Tool for Software Development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 667–676). Florence, Italy: ACM. http://doi.org/10.1145/1357054.1357160

Paul, M. J., & Dredze, M. (2011). You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)* (pp. 265–272).

Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*.

Plaisant, C. (2004). The Challenge of Information Visualization Evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI)*. ACM. http://doi.org/10.1145/989863.989880

Plaisant, C., Grinstein, G., & Scholtz, J. (2009). Visual-Analytics Evaluation. *IEEE Computer Graphics and Applications*, (May/June 2009), 16–17.

Plutchik, R. (1991). *The Emotions*. Lanham, MD: University Press of America.

Plutchik, R. (2001). The Nature of Emotions. *American Scientist*, *89*(4), 344–350.

Poon, S. S., Thomas, R. C., Aragon, C. R., & Lee, B. (2008). Context-Linked Virtual Assistants for Distributed Teams: An Astrophysics Case Study. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW)*. San Diego, CA: ACM Press. http://doi.org/10.1145/1460563.1460623

Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011). Identifying Health-Related Topics on Twitter: An Exploration of Tobacco-Related Tweets as a Test Topic. In J. Salerno, S. J. Yang, D. Nau, & S.-K. Chai (Eds.), *Social Computing, Behavioral-Cultural Modeling and Prediction* (pp. 18–25). Springer Berlin Heidelberg.

Purver, M., & Battersby, S. (2012). Experimenting with Distant Supervision for Emotion Classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 482–491). Stroudsburg, PA, USA: Association for Computational Linguistics.

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). Tracking "Gross Community Happiness" from Tweets. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)* (pp. 965–968). Seattle, WA: ACM. http://doi.org/10.1145/2145204.2145347

Quinn, A. J., & Bederson, B. B. (2011). Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1403–1412). Vancouver, BC: ACM. http://doi.org/10.1145/1978942.1979148

Raghavan, H., Madani, O., & Jones, R. (2005). InterActive Feature Selection. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (pp. 841–846). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rahimi, A., Vu, D., Cohn, T., & Baldwin, T. (2015). Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*. Denver, CO: ACL. Retrieved from http://aclweb.org/anthology/N/N15/N15-1153.pdf

Ranganathan, A., Campbell, R. H., Ravi, A., & Mahajan, A. (2002). ConChat: a context-aware chat program. *IEEE Pervasive Computing*, *1*(3), 51–57. http://doi.org/10.1109/MPRV.2002.1037722

Richards, L. (2014). *Handling qualitative data: A practical guide*. SAGE Publications, Inc.

Risch, J., Kao, A., Poteet, S. R., & Wu, Y.-J. J. (2008). Text visualization for visual text analytics. *Visual Data Mining*, 154–171. http://doi.org/10.1007/978-3-540-71080-6_11

Rogerson, C., & Scott, E. (2010). The Fear Factor: How It Affects Students Learning to Program in a Tertiary Environment. *Journal of Information Technology Education: Research*, *9*(1), 147–171.

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal on Computer-Supported Collaborative Learning*, *3*(3), 237–271. http://doi.org/10.1007/s11412-007-9034-0

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The Cost Structure of Sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (pp. 269–276). Amsterdam, Netherlands: ACM. http://doi.org/10.1145/169059.169209

Russ, S. W. (1993). *Affect and Creativity: The Role of Affect and Play in the Creative Process*. Routledge.

Sawyer, S., Kaziunas, E., & Øesterlund, C. (2012). Social Scientists and Cyberinfrastructure: Insights from a Document Perspective. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)* (pp. 931–934). Seattle, WA: ACM. http://doi.org/10.1145/2145204.2145342

Schmidt, K. (2000). The critical role of workplace studies in CSCW. In P. Luff, J. Hindmarsh, & C. Heath (Eds.), *Workplace studies: Recovering work practice and informing system design*. Cambridge University Press.

Scott, S., & Matwin, S. (1999). Feature engineering for text classification. In *Proceedings of the 16th International Conference on Machine Learning (ICML)* (Vol. 99, pp. 379–388).

Scott, T. J., Kuksenok, K., Perry, D., Brooks, M., Anicello, O., & Aragon, C. R. (2012). Adapting Grounded Theory to Construct a Taxonomy of Affect in Collaborative Online Chat. In *Proceedings of the 30th ACM International Conference on Design of Communication (SIGDOC)*. Seattle, WA: ACM. http://doi.org/10.1145/2379057.2379096

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1), 1–47. http://doi.org/10.1145/505282.505283

Sedlmair, M., Isenberg, P., Baur, D., Mauerer, M., Pigorsch, C., & Butz, A. (2011). Cardiogram: Visual Analytics for Automotive Engineers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1727–1736). Vancouver, BC: ACM. http://doi.org/10.1145/1978942.1979194

Seidman, I. (2005). *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences* (Third Edition). New York: Teachers College Press.

Simard, P., Chickering, D., Lakshmiratan, A., Charles, D., Bottou, L., Suarez, C. G. J., … Suh, J. (2014). ICE: Enabling Non-Experts to Build Models Interactively for Large-Scale Lopsided Problems. *arXiv:1409.4814 [cs]*. Retrieved from http://arxiv.org/abs/1409.4814

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online*, *18*(3). Retrieved from http://econpapers.repec.org/article/srosrosro/2012-178-2.htm

Starbird, K., Dailey, D., Walker, A. H., Leschine, T. M., Pavia, R., & Bostrom, A. (2015). Social Media, Public Participation, and the 2010 BP Deepwater Horizon Oil Spill. *Human and Ecological Risk Assessment*, *21*(3), 605–630. http://doi.org/10.1080/10807039.2014.947866

Starbird, K., & Palen, L. (2012). (How) Will the Revolution Be Retweeted?: Information Diffusion and the 2011 Egyptian Uprising. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)* (pp. 7–16). Seattle, WA: ACM. http://doi.org/10.1145/2145204.2145212

St John, W., & Johnson, P. (2000). The Pros and Cons of Data Analysis Software for Qualitative Research. *Journal of Nursing Scholarship*, *32*(4), 393–397. http://doi.org/10.1111/j.1547-5069.2000.00393.x

Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Retrieved from http://www.lrec-conf.org/proceedings/lrec2004/pdf/369.pdf

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park: SAGE Publications, Inc.

Stumpf, S., Rajaram, V., Li, L., Burnett, M., Dietterich, T., Sullivan, E., … Herlocker, J. (2007). Toward harnessing user feedback for machine learning. In *Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI)*. Honolulu, HI: ACM. http://doi.org/10.1145/1216295.1216316

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307. http://doi.org/10.1162/COLI_a_00049

Tak, S., Westendorp, P., & van Rooij, I. (2013). Satisficing and the Use of Keyboard Shortcuts: Being Good Enough Is Enough? *Interacting with Computers*, *25*(5), 404–416. http://doi.org/10.1093/iwc/iwt016

Talbot, J., Lee, B., Kapoor, A., & Tan, D. S. (2009). EnsembleMatrix: Interactive Visualization to Support Machine Learning with Multiple Classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1283–1292). Boston, MA: ACM. http://doi.org/10.1145/1518701.1518895

Tarasov, A., Delany, S. J., & Cullen, C. (2010). Using crowdsourcing for labelling emotional speech assets. In *W3C Workshop on Emotion ML*. Paris, France. Retrieved from http://arrow.dit.ie/dmccon/49/

Tastes, Ties, and Time: Facebook data release. (2008, September 25). Retrieved September 11, 2015, from https://cyber.law.harvard.edu/node/94446

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. http://doi.org/10.1177/0261927X09351676

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558. http://doi.org/10.1002/asi.21416

Thomas, J. J., & Cook, K. A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alamitos, CA: IEEE.

Toomim, M., Kriplean, T., Pörtner, C., & Landay, J. (2011). Utility of human-computer interactions: Toward a science of preference measurement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 2275–2284). Vancouver, BC: ACM. http://doi.org/10.1145/1978942.1979277

Torkildson, M. K. (2013). Visualizing the Performance of Classification Algorithms with Additional Re-annotated Data. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (pp. 2767–2772). Paris, France: ACM. http://doi.org/10.1145/2468356.2479507

Tory, M., & Staub-French, S. (2008). Qualitative Analysis of Visualization: A Building Design Field Study. In *Proceedings of the CHI 2008 Workshop on BEyond Time and Errors: Novel evaLuation Methods for Information Visualization (BELIV)*. Florence, Italy: ACM. http://doi.org/10.1145/1377966.1377975

Vidich, A. J., & Lyman, S. M. (1993). Qualitative methods: Their history in sociology and anthropology. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE Handbook of Qualitative Research* (First Edition). SAGE Publications, Inc.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1079–1088). Atlanta, GA: ACM. http://doi.org/10.1145/1753326.1753486

Wang, L., & Oard, D. W. (2009). Context-based Message Expansion for Disentanglement of Interleaved Text Conversations. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)* (pp. 200–208). Boulder, CO: Association for Computational Linguistics.

Ware, M., Frank, E., Holmes, G., Hall, M., & Witten, I. H. (2001). Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, *55*(3), 281–292. http://doi.org/10.1006/ijhc.2001.0499

Wattenberg, M. (2005). Baby names, visualization, and social data analysis. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*. Minneapolis, MN: IEEE. http://doi.org/10.1109/INFVIS.2005.1532122

Wattenberg, M., & Kriss, J. (2006). Designing for social data analysis. *IEEE Transactions on Visualization and Computer Graphics*, *12*(4), 549–557. http://doi.org/10.1109/TVCG.2006.65

Weigend, A. (2009, May 20). The Social Data Revolution(s). Retrieved from http://blogs.hbr.org/2009/05/the-social-data-revolution/

Weiss, R. S. (1995). *Learning From Strangers: The Art and Method of Qualitative Interview Studies*. New York: Free Press.

Willett, W., Heer, J., Hellerstein, J., & Agrawala, M. (2011). CommentSpace: Structured Support for Collaborative Visual Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3131–3140). Vancouver, BC: ACM. http://doi.org/10.1145/1978942.1979407

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques* (Third Edition). Morgan Kaufmann.

Zafarani, R., Cole, W. D., & Liu, H. (2010). Sentiment Propagation in Social Networks : A Case Study in LiveJournal. In *International Social Computing, Behavioral Modeling and Prediction Conference* (pp. 413–420).

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring International and Internal Migration Patterns from Twitter Data. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)* (pp. 439–444). Geneva, Switzerland. http://doi.org/10.1145/2567948.2576930

Zagheni, E., & Weber, I. (2012). You Are Where You e-Mail: Using e-Mail Data to Estimate International Migration Rates. In *Proceedings of the 4th Annual ACM Web Science Conference (WebSci)* (pp. 348–351). Evanston, IL: ACM. http://doi.org/10.1145/2380718.2380764

Zimmer, M. (2010). "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313–325. http://doi.org/10.1007/s10676-010-9227-5

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research Through Design As a Method for Interaction Design Research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 493–502). San Jose, CA: ACM. http://doi.org/10.1145/1240624.1240704