1

TEXT MINING AND TEXT ANALYSIS

LEARNING OBJECTIVES

The goals of Chapter 1 are to help you to do the following:

1. Familiarize yourself with a variety of research projects accomplished using **text mining** tools.
2. Address different research questions using text mining tools.
3. Differentiate between text mining and **text analysis** methodologies.
4. Compare major theoretical and methodological approaches to both text mining and text analysis.

INTRODUCTION

Text mining is an exciting field that encompasses new research methods and software tools that are being used across academia as well as by companies and government agencies. Researchers today are using text mining tools in ambitious projects to attempt to predict everything from the direction of stock markets (Bollen, Mao, & Zeng, 2011) to the occurrence of political protests (Kallus, 2014). Text mining is also commonly used in marketing research and many other business applications as well as in government and defense work.

Over the past few years, text mining has started to catch on in the social sciences, in academic disciplines as diverse as anthropology (Acerbi, Lampos, Garnett, & Bentley, 2013; Marwick, 2013), communications (Lazard, Scheinfeld, Bernhardt, Wilcox, & Suran, 2015), economics (Levenberg, Pulman, Moilanen, Simpson, & Roberts, 2014), education (Evison, 2013), political science (Eshbaugh-Soha, 2010; Grimmer & Stewart, 2013), psychology (Colley & Neal, 2012; Schmitt, 2005), and sociology (Bail, 2012; Heritage & Raymond, 2005; Mische, 2014). Before social scientists began to adapt text mining tools to use in their research, they spent decades studying transcribed interviews, newspaper articles, speeches, and other forms of textual data, and they developed sophisticated text analysis methods that we review in the chapters in Part IV. So while text mining is a relatively new interdisciplinary field based in computer science, text analysis methods have a long history in the social sciences (see Roberts, 1997).

Text mining processes typically include information retrieval (methods for acquiring texts) and applications of advanced statistical methods and **natural language processing (NLP)** such as part-of-speech tagging and syntactic parsing. Text mining also often involves named entity recognition (NER), which is the use of statistical techniques to identify named text features such as people, organizations, and place names; **disambiguation**, which is the use of contextual clues to decide where words refer to one or another of their multiple meanings; and **sentiment analysis**, which involves discerning subjective material and extracting attitudinal information such as sentiment, opinion, mood, and emotion. These techniques are covered in Parts III and V of this book. Text mining also involves more basic techniques for acquiring and processing data. These techniques include tools for **web scraping** and **web crawling**, for making use of dictionaries and other lexical resources, and for processing texts and relating words to texts. These techniques are covered in Parts II and III.

RESEARCH IN THE SPOTLIGHT

Predicting the Stock Market With Twitter

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

The computer scientists Bollen, Mao, and Zeng asked whether societies can experience mood states that affect their collective decision making, and by extension whether the public mood is correlated or even predictive of economic indicators. Applying sentiment analysis (see Chapter 14) to large-scale Twitter feeds, Bollen and colleagues investigated whether measurements of collective mood states are correlated to the value of the Dow Jones Industrial Average over time. They analyzed the text content of daily Twitter feeds using OpinionFinder, which measures positive versus negative mood and Google Profile of Mood States to measure mood in terms of six dimensions (calm, alert, sure, vital, kind, and happy). They also investigated the hypothesis that public mood states are predictive of changes in Dow Jones Industrial Average closing values, finding that the accuracy of stock market predictions can be significantly improved by the inclusion of some specific public mood dimensions but not others.

Specialized software used:

OpinionFinder

http://mpqa.cs.pitt.edu/opinionfinder

Text analysis involves systematic analysis of word use patterns in texts and typically combines formal statistical methods and less formal, more humanistic interpretive techniques. Text analysis arguably originated as early as the 1200s with the Dominican friar Hugh of Saint-Cher and his team of several hundred fellow friars who created the first biblical **concordance**, or cross-listing of terms and concepts in the Bible. There is also evidence of European inquisitorial church studies of newspapers in the late 1600s, and the first well-documented quantitative text analysis was performed in Sweden in the 1700s when the Swedish state church analyzed the symbology and ideological content of popular hymns that appeared to challenge church orthodoxy (Krippendorff, 2013, pp. 10–11). The field of text analysis expanded rapidly in the 20th century as researchers in the social sciences and humanities developed a broad spectrum of techniques for analyzing texts, including methods that relied heavily on human interpretation of texts as well as formal statistical methods. Systematic quantitative analysis of newspapers was performed in the late 1800s and early 1900s by researchers including Speed (1893), who showed that in the late 1800s New York newspapers had decreased their coverage of literary, scientific, and religious matters in favor of sports, gossip, and scandals. Similar text analysis studies were performed by Wilcox (1900), Fenton (1911), and White (1924), all of whom quantified newspaper space devoted to different categories of news. In the 1920s through 1940s, Lasswell and his colleagues conducted breakthrough **content analysis** studies of political messages and propaganda (e.g., Lasswell, 1927). Lasswell's work inspired large-scale content analysis projects including the **General Inquirer project** at Harvard, which is a lexicon attaching syntactic, semantic, and pragmatic information to part-of-speech tagged words (Stone, Dunphry, Smith, & Ogilvie, 1966).

While text mining's roots are in computer science and the roots of text analysis are in the social sciences and humanities, today, as we will see throughout this textbook, the two fields are converging. Social scientists and humanities scholars are adapting text mining tools for their research projects, while text mining specialists are investigating the kinds of social phenomena (e.g., political protests and other forms of collective behavior) that have traditionally been studied within the social sciences.

SIX APPROACHES TO TEXT ANALYSIS

The field of text mining is divided mainly in terms of different methodologies, while the field of text analysis can be divided into several different approaches that are each based on a different way of theorizing language use. Before discussing some of the special challenges associated with using online data for social science research, next we review six of the most prominent approaches to text analysis. As we will see, many researchers who work with these approaches are finding ways to make use of the new text mining methodologies and tools that are covered in Parts II, III, and V. These approaches include **conversation analysis**, analysis of **discourse positions**, **critical discourse analysis (CDA)**, content analysis, **Foucauldian analysis**, and analysis of texts as social information. These approaches use different logical strategies and are based on different theoretical foundations and philosophical assumptions (discussed in Chapter 4). They also operate at different levels of analysis (micro, meso, and macro) and employ different selection and sampling strategies (see Chapter 5).

Conversation Analysis

Conversation analysts study everyday conversations in terms of how people negotiate the meaning of the conversation in which they are participating and the larger discourse of which the conversation is a part. Conversation analysts focus not only on what is said in daily conversations but also on how people use language pragmatically to define the situations in which they find themselves. These processes go mostly unnoticed until there is disagreement as to the meaning of a particular situation. An example of conversation analysis is the educational researcher Evison's (2013) study of "academic talk," which used corpus linguistic techniques (see Appendix F) on both a corpus of 250,000 words of spoken academic discourse and a benchmark corpus of casual conversation to explore conversational turn openings. The corpus of academic discourse included 13,337 turns taken by tutors and students in a range of social interactions. In seeking to better understand the unique language of academia and of specific academic disciplines, Evison identified six items that have a particularly strong affinity with the turn-opening position (*mhm, mm, yes, laughter, oh, no*) as key characteristics of academic talk.

Further examples of conversation analysis research include studies of conversation in educational settings by O'Keefe and Walsh (2012); in health care settings by Heath and Luff (2000), Heritage and Raymond (2005), and Silverman (2016); and in online environments among Wikipedia editors by Danescu-Niculescu-Mizil, Lee, Pang, and Kleinberg (2012). O'Keefe and Walsh's 2012 study combined corpus linguistics and conversation analysis methodologies to analyze higher education small-group teaching sessions. Their data are from a 1-million-word corpus, the Limerick–Belfast Corpus of Academic Spoken English (LIBEL CASE). Danescu-Niculescu-Mizil and colleagues (2012) analyzed signals manifested in language in order to learn about roles, status, and other aspects of groups' interactional dynamics. In their study of Wikipedians and of arguments before the U.S. Supreme Court, they showed that in group discussions, power differentials between participants are subtly revealed by the degree to which one individual immediately echoes the linguistic style of the person to whom they are responding. They proposed an analysis framework based on linguistic coordination that can be used to shed light on power relationships and that works consistently across multiple types of power, including more static forms of power based on status differences and more situational forms in which one individual experiences a type of dependence on another.

Hakimnia and her colleagues' (2015) conversation analysis of transcripts of calls to a telenursing site in Sweden used a comparative research design (see Chapter 5). The study's goal was to analyze callers' reasons for calling and the outcome of the calls in terms of whether men and women received different kinds of referrals. The researchers chose to randomly sample 800 calls from a corpus of over 5,000 total calls that had been recorded at a telenursing site in Sweden over a period of 11 months. Callers were informed about the study in a prerecorded message and consented to participate, while the nurses were informed verbally about the study. The first step in the analysis of the final sample of 800 calls was to create a matrix (see Chapter 5 and Appendices C and D), including information on each caller's gender, age, fluency or nonfluency in Swedish as well as the outcome of the call (whether callers were referred to a general practitioner). The researchers found that men, and especially fathers, received more referrals to general practitioners than did women. The most common caller was a woman fluent in Swedish (64%), and the least likely caller was a man nonfluent in Swedish (3%). All in all, 70% of the callers were women. When the calls concerned children, 78% of the callers were female. Based on these results, the researchers concluded that it is important that telenursing not become a "feminine" activity, only suitable for young callers fluent in Swedish. Given the telenurses' gatekeeping role, there is a risk that differences on this first level of health care could be reproduced throughout the whole health care system.

Analysis of Discourse Positions

Analyzing discourse positions is an approach to text analysis that allows researchers to reconstruct communicative interactions through which texts are produced and in this way gain a better understanding of their meaning from their author's viewpoint. Discourse positions are understood as typical discursive roles that people adopt in their everyday communication practices, and the analysis of discourse positions is a way of linking texts to the social spaces in which they have emerged. An example of contemporary discourse position research is Bamberg's (2004) study of the "small stories" told by adolescents and postadolescents about their identities. Bamberg's 2004 study is informed by theories of human development and of narrative (see Chapter 10). His texts are excerpts of transcriptions from a group discussion among five 15-year-old boys telling a story about a female student they all know. The group discussion was conducted in the presence of an adult moderator, but the data were collected as part of a larger project in which Bamberg and his colleagues collected journal entries and transcribed oral accounts from 10-, 12-, and 15-year-old boys in one-on-one interviews and group discussions. Although the interviews and groups discussions were open-ended, they all focused on the same list of topics, including friends and friendships, girls, the boys' feelings and sense of self, and their ideas about adulthood and future orientation. Bamberg and his team analyzed the transcripts line by line, coding instances of the boys positioning themselves relative to each other and to characters in their stories.

Edley and Wetherell's (1997, 2001; Wetherell & Edley, 1999) studies of masculine identity formation are similar to Bamberg's study in that they also focus on stories people tell themselves and others in ordinary everyday conversations. Edley and Wetherell studied a corpus of men's talk on feminism and feminists to identify patterns and regularities in their accounts of feminism and in the organization of their rhetoric. Their samples of men included a sample of white, middle-class 17- to 18-year-old school students and a sample of 60 interviews with a more diverse sample of older men aged 20 to 64. The researchers identified two "interpretative repertoires of feminism and feminists," which set up a "Jekyll and Hyde" binary and "positioned feminism along with feminists very differently as reasonable versus extreme" (Edley & Wetherell, 2001, p. 439).

In the end, analysis of discourse positions is for the most part a qualitative approach to text analysis that relies almost entirely on human interpretation of texts (see Hewson, 2014). Appendix D includes a list of contemporary qualitative data analysis software (QDAS) packages that can be used to organize and code the kinds of text corpora analyzed by Bamberg, Edley, Wetherell, and other researchers working in this tradition.

Critical Discourse Analysis

CDA involves seeking the presence of features from other discourses in the text or discourse to be analyzed. CDA is based on Fairclough's (1995) concept of "intertextuality," which is the idea that people appropriate from discourses circulating in their social space whenever they speak or write. In CDA, ordinary everyday speaking and writing are understood to involve selecting and combining elements from dominant discourses.

While the term *discourse* generally refers to all practices of writing and talking, in CDA discourses are understood as ways of writing and talking that "rule out" and "rule in" ways of constructing knowledge about topics. In other words, discourses "do not just describe things; they do things" (Potter & Wetherell, 1987, p. 6) through the way they make sense of the world for its inhabitants (Fairclough, 1992; van Dijk, 1993).

Discourses cannot be studied directly but can be explored by examining the texts that constitute them (Fairclough, 1992; Parker, 1992). In this way, texts can be analyzed as fragments of discourses that reflect and project ideological domination by powerful groups in society. But texts can also be considered a potential mechanism of liberation when they are produced by the critical analyst who reveals mechanisms of ideological domination in them in an attempt to overcome or eliminate them.

Although CDA has generally employed strictly interpretive methods, use of quantitative and statistical techniques is not a novel practice (Krishnamurthy, 1996; Stubbs, 1994), and the use of software to create, manage, and analyze large collections of texts appears to be increasingly popular (Baker et al., 2008; Koller & Mautner, 2004; O'Halloran & Coffin, 2004).

A 2014 study by Bednarek and Caple exemplifies the use of statistical techniques in CDA. Bednarek and Caple introduced the concept of "news values" to CDA of news media and illustrated their approach with two case studies using the same collection of British news discourse. Their texts included 100 news stories (about 70,000 words total) from 2003 covering 10 topics from 10 different national newspapers, including five quality papers and five tabloids. The analysis proceeded through analysis of word frequency of the top 100 most frequently used words and two-word clusters (bigrams), focusing on words that represent news values such as *eliteness, superlativeness, proximity, negativity, timeliness, personalization*, and *novelty*. The authors concluded that their case studies demonstrated that corpus linguistic techniques (see Appendix F) can identify discursive devices that are repeatedly used in news discourse to construct and perpetuate an ideology of newsworthiness.

In another CDA study, Baker and his colleagues (2008) analyzed a 140-million-word corpus of British news articles about refugees, asylum seekers, immigrants, and migrants. They used collocation and concordance analysis (see Appendix F) to identify common categories of representation of refugees, asylum seekers, immigrants, and migrants. They also discussed how collocation and concordance analysis can be used to direct researchers to representative texts in order to carry out qualitative analysis.

RESEARCH IN THE SPOTLIGHT

Combining Critical Discourse Analysis and Corpus Linguistics

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., Mcenery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273–306.

In this critical discourse analysis (CDA) study, the linguist Baker and his colleagues analyzed a 140-million-word corpus of British news articles about refugees, asylum seekers, immigrants, and migrants. The authors used collocation and concordance analysis (see Appendix F) to identify common categories of representations of the four groups. The authors also discuss how collocation and concordance analysis can be used to direct researchers to representative texts in order to carry out qualitative analysis.

Specialized software used:

WordSmith

www.lexically.net/wordsmith

Content Analysis

Content analysis adopts a quantitative, scientific approach to text analysis. Unlike CDA, content analysis is generally focused on texts themselves rather than texts' relations to their social and historical contexts. One of the classic definitions of content analysis defines it as "a research technique for the objective, systematic-quantitative description of the manifest content of communication" (Berelson, 1952, p. 18). At a practical level, content analysis involves the development of a coding frame that is applied to textual data. It mainly consists of breaking down texts into pertinent units of information in order to permit subsequent coding and categorization.

Krippendorff's (2013) classic textbook *Content Analysis* is the standard reference for work in this area. Many of the research design principles and sampling techniques covered in Chapter 5 of this textbook are shared with content analysis, although Krippendorff's book goes into much greater detail on statistical sampling of texts and units of texts, as well as on statistical tests of interrater reliability.

Foucauldian Analysis

The philosopher and historian Foucault (1973) developed an influential conceptualization of intertextuality that differs significantly from Fairclough's conceptualization in CDA. Rather than identifying the influence of external discourses within a text, for Foucault the meaning of a text emerges in reference to discourses with which it engages in dialogue. These engagements may be explicit or, more often, implicit. In Foucauldian intertextual analysis, the analyst must ask each text about its presuppositions and with which discourses it dialogues. The meaning of a text therefore derives from its similarities and differences with respect to other texts and discourses and from implicit presuppositions within the text that can be recognized by historically informed close reading.

Foucauldian analysis of texts is performed in many theoretical and applied research fields. For instance, a number of studies have used Foucauldian intertextual analysis to analyze forestry policy (see Winkel, 2012, for an overview). Researchers working in Europe (e.g., Berglund, 2001; Franklin, 2002; Van Herzele, 2006), North America, and developing countries (e.g., Asher & Ojeda, 2009; Mathews, 2005) have used Foucauldian analysis to study policy discourses regarding forest management, forest fires, and corporate responsibility.

Another example of Foucauldian intertextual analysis is a sophisticated study of the professional identities of nurses by Bell, Campbell, and Goldberg (2015). Bell and colleagues argued that nurses' professional identities should be understood in relation to the identities of other occupational categories within the health care field. The authors collected their data from PubMed, a medical research database. Using PubMed's own user interface, the authors acquired the abstracts for research papers that used the terms *service* or *services* in the abstract or key words for a period from 1986 to 2013. The downloaded abstracts were added to an SQLite database, which was used to generate comma-separated values (CSV) files with abstracts organized into 3-year periods. The authors then spent approximately 6 weeks of full-time work, manually checking the data for duplicates and other errors. The final sample included over 230,000 abstracts. Bell and colleagues then used the text analysis package Leximancer (see Appendix C) to calculate frequency and co-occurrence statistics for all concepts in the abstracts (see also Appendix F). Leximancer also produced concept maps (see Appendix G) to visually represent the relationships between concepts. The authors further cleaned their data after viewing these initial concept maps and finding a number of irrelevant terms and then used Leximancer to analyze the concept of nursing in terms of its co-occurrence with other concepts.

Analysis of Texts as Social Information

Another category of text analysis treats texts as reflections of the practical knowledge of their authors. This type of analysis is prevalent in grounded theory studies (see Chapter 4) as well as in applied studies of expert discourses. Interest in the informative analysis of texts is due in part to its practical value, because user-generated texts can potentially provide analysts with reliable information about social reality. Naturally, the quality of information about social reality that is contained in texts varies according to the level of knowledge of each individual who has participated in the creation of the text, and the information that subjects provide is partial insofar as it is filtered by their own particular point of view.

An example of analysis of texts as social information is a 2012 psychological study by Colley and Neal on the topic of organizational safety. Starting with small representative samples of upper managers, supervisors, and workers in an Australian freight and passenger rail company, Colley and Neal conducted open-ended interviews with members of the three groups. These were transcribed and analyzed using Leximancer (see Appendix C) for map analysis (see also Appendix G). Comparing the concept maps produced for the three groups revealed significant differences between the "safety climate schema" of upper managers, supervisors, and workers.

CHALLENGES AND LIMITATIONS OF USING ONLINE DATA

Having introduced text mining and text analysis, in this section we review some lessons that have been learned from other fields about how best to adapt social science research methods to data from online environments. This section is short but critically important for students who plan to perform research with data taken from social media platforms and websites.

Methodologies such as text mining that analyze data from digital environments offer potential cost- and time-efficiency advantages over older methods (Hewson & Laurent, 2012; Hewson, Yule, Laurent, & Vogel, 2003), as the Internet provides ready access to a potentially vast, geographically diverse participant pool. The speed and global reach of the Internet can facilitate cross-cultural research projects that would otherwise be prohibitively expensive. It also allows for the emergence of patterns of social interactions, which are elaborate in terms of their richness of communication exchange but where levels of anonymity and privacy can be high. The Internet's unique combination of digital archiving technologies and users' perceptions of anonymity and privacy may reduce social desirability effects (where research participants knowingly or unknowingly attempt to provide researchers with socially acceptable and desirable, rather than accurate, information). The unique attributes of Internet-based technologies may also reduce biases resulting from the perception of attributes such as race, ethnicity, and sex or gender, promoting greater candor. The convenience of these technologies can also empower research participants by allowing them to take part in study procedures that fit their schedules and can be performed within their own spaces such as at home or in a familiar work environment.

While Internet-based research has many advantages (see Hewson, Vogel, & Laurent, 2015), Internet-based data have a number of serious drawbacks for social science research. One major disadvantage is the potentially biased nature of Internet-accessed data samples. **Sample bias** is one of the most fundamental and difficult to manage challenges associated with Internet-mediated research (see Chapter 5). Second, as compared to offline methods, Internet-based data are often characterized by reduced levels of researcher control. This lack of control arises mainly from technical issues, such as users' different hardware and software configurations and network traffic performance. Research participants working with different hardware platforms, operating systems, and browsers may experience social media services and online surveys very differently, and it is often extremely difficult for researchers to fully appreciate differences in participants' experiences. In addition, hardware and software failures may lead to unpredicted effects, which may cause problems. Because of the lack of researcher presence, in Internet-based research there is often a lack of researcher control over and knowledge of variations in participants' behaviors and the participation context. This may cause problems related to the extent to which researchers can gauge participants' intentions and levels of sincerity and honesty during a study, as researchers lack nonverbal cues to evaluate participants compared with face-to-face communication.

Despite these weaknesses, scholars have long recognized digital technologies' potential as research tools. While social researchers have occasionally developed brand-new Internet-based methodologies, they have also adapted preexisting research methods for use with evolving digital technology. Because a number of broadly applicable lessons have been learned from these adaptation processes, in the remainder of this chapter we briefly review some of the most widely used social science research methods that have been adapted to Internet-related communication technologies and some of the lessons learned from each. We discuss offline and online approaches to *social surveys*, *ethnography*, and *archival research* but do not cover online focus groups (Krueger & Casey, 2014) or experiments (Birnbaum, 2000). While focus groups and experiments are both important and widely used research methods, we have found that the lessons learned from developing online versions of these methods are less applicable to text mining than lessons learned from the former three.

Social Surveys

Social surveys are one of the most commonly used methods in the social sciences, and researchers have been working with online versions of surveys since the 1990s. Traditional telephone and paper surveys tend to be costly, even when using relatively small samples, and the costs of a traditional large-scale survey using mailed questionnaires can be enormous. Although the costs of online survey creation software and web survey services vary widely, by eliminating the need for paper, postage, and data entry costs, online surveys are generally less expensive than their paper- and telephone-based equivalents (Couper, 2000; Ilieva, Baron, & Healey, 2002; Yun & Trumbo, 2000). Online surveys can also save researchers time by allowing them to quickly reach thousands of people despite possibly being separated by great geographic distances (Garton, Haythornthwaite, & Wellman, 2007). With an online survey, a researcher can quickly gain access to large populations by posting invitations to participate in the survey to newsgroups, chat rooms, and message boards. In addition to their cost and time savings and overall convenience, another advantage of online surveys is that they exploit the ability of the Internet to provide access to groups and individuals who would be difficult, if not impossible, to reach otherwise (Garton et al., 1997).

While online surveys have significant advantages over paper- and phone-based surveys, they bring with them new challenges in terms of applying traditional survey research methods to the study of online behavior. Online survey researchers often encounter problems regarding sampling, because relatively little may be known about the characteristics of people in online communities aside from some basic demographic variables, and even this information may be questionable (Walejko, 2009). While attractive, features of online surveys themselves, such as multimedia, and of online survey services, such as use of company e-mail lists to generate samples, can affect the quality of the data they produce in a variety of ways.

The process of adapting social surveys to online environments offers a cautionary lesson for text mining researchers. The issue of user demographics casts a shadow over online survey research just as it does for text mining, because in online environments it is very difficult for researchers to make valid inferences about their populations of interest. The best practice for both methodologies is for researchers to carefully plan and then explain in precise detail their sampling strategies (see Chapter 5).

Ethnography

In the 1990s, researchers began to adapt ethnographic methods designed to study geographically situated communities to online environments which are characterized by relationships that are technologically mediated rather than immediate (Salmons, 2014). The result is **virtual ethnography** (Hine, 2000) or **netnography** (Kozinets, 2009), which is the ethnographic study of people interacting in a wide range of online environments. Kozinets, a netnography pioneer, argues that successful netnography requires researchers to acknowledge the unique characteristics of these environments and to effect a “radical shift” from offline ethnography, which observes people, to a mode of analysis that involves recontextualizing conversational acts (Kozinets, 2002, p. 64). Because netnography provides more limited access to fixed demographic markers than does ethnography, the identities of discussants are much more difficult to discern. Yet netnographers must learn as much as possible about the forums, groups, and individuals they seek to understand. Unlike in traditional ethnographies, in the identification of relevant communities, online search engines have proven invaluable to the task of learning about research populations (Kozinets, 2002, p. 63).

Just as the quality of social survey research depends on sampling, netnography requires careful case selection (see Chapter 5). Netnographers must begin with specific research questions and then identify online forums appropriate to these questions (Kozinets, 2009, p. 89).

Netnography's lessons for text mining and analysis are straightforward. Leading researchers have shown that for netnography to be successful, researchers must acknowledge the unique characteristics of online environments, recognize the importance of developing and explaining their data selection strategy, and learn as much as they possibly can about their populations of interest. All three lessons apply to text mining research that analyzes user-generated data mined from online sources.

Historical Research Methods

Archival research methods are among the oldest methods in the social sciences. The founding fathers of sociology—Marx, Weber, and Durkheim—all did historical scholarship based on archival research, and today, archival research methods are widely used by historians, political scientists, and sociologists.

Historical researchers have adapted digital technology to archival research in two waves. The first occurred in the 1950s and 1960s when, in the early years of accessible computers, historians taught themselves statistical methods and programming languages. Adopting quantitative methods developed in sociology and political science, during this period historians made lasting contributions in the areas of "social mobility, political identification, family formation, patterns of crime, economic growth, and the consequences of ethnic identity" (Ayers, 1999). Unfortunately, however, that quantitative social science history collapsed suddenly, the victim of its own inflated claims, limited method and machinery, and changing academic fashion. By the mid-80s, history, along with many of the humanities and social sciences, had taken the linguistic turn. Rather than SPSS guides and codebooks, innovative historians carried books of French philosophy and German literary interpretation. The social science of choice shifted from sociology to anthropology; texts replaced tables. A new generation defined itself in opposition to social scientific methods just as energetically as an earlier generation had seen in those methods the best means of writing a truly democratic history. The first computer revolution largely failed (Ayers, 1999).

Beginning in the 1980s, historians and historically minded social scientists began to reengage with digital technologies. While today historical researchers use digital technologies at every stage of the research process, from professional communication to multimedia presentations, **digital archives** have had perhaps the most profound influence on the practice of historical research. Universities, research institutes, and private companies have digitized and created accessible archives of massive volumes of historical documents.

Historians recognize that these archives offer tremendous advantages in terms of the capacity, flexibility, accessibility, flexibility, diversity, manipulability, and interactivity of research (Cohen & Rosenzweig, 2005). However, digital research archives also pose dangers in terms of the quality, durability, and readability of stored data. There is also a potential for inaccessibility and monopoly and also for digital archives to encourage researcher passivity (Cohen & Rosenzweig, 2005).

There are lessons to be learned from digital history for text mining and text analysis, particularly from the sudden collapse of the digital history movement of the 1950s and 1960s. In light of the failure of that movement, it is imperative that social scientists working with text mining tools recognize the limitations of their chosen methods and not make imperious or inflated claims about these tools' revolutionary potential. Like all social science methods, text mining methods have benefits and drawbacks that must be recognized from the start and given consideration in every phase of the research process. And text mining researchers should be aware of historians' concerns about the quality of data stored in digital archives and the possibility for digital archives to encourage researcher passivity in the data gathering phase of research.

Conclusion

This chapter has introduced text mining and text analysis methodologies, provided an overview of the major approaches to text analysis, and discussed some of the risks associated with analyzing data from online sources. Despite these risks, social and computer scientists are developing new text mining and text analysis tools to address a broad spectrum of applied and theoretical research questions, in academia as well as in the private and public sectors.

In the chapters that follow, you will learn how to find data online (Chapters 2 and 6), and you will learn about some of the ethical (Chapter 3) and philosophical and logical (Chapter 4) dimensions of text mining research. In Chapter 5, you will learn how to design your own social science research project. Parts II, IV, and V review specific text mining techniques for collecting and analyzing data, and Chapter 17 in Part VI provides guidance for writing and reporting your own research.

Key Terms (see Glossary)

Concordance 5
Content analysis 5
Conversation analysis 6
Critical discourse analysis (CDA) 6
Digital archives 15
Disambiguation 4
Discourse positions 6
Foucauldian analysis 6

General Inquirer project 5
Natural language processing (NLP) 4
Netnography 14
Sample bias 12
Sentiment analysis 4
Text analysis 3
Text mining 3
Virtual ethnography 14
Web crawling 4
Web scraping 4

Highlights

- Text mining processes include methods for acquiring digital texts and analyzing them with NLP and advanced statistical methods.
- Text mining is used in many academic and applied fields to analyze and predict public opinion and collective behavior.
- Text analysis began with analysis of religious texts in the Middle Ages and was developed by social scientists starting in the early 20th century.
- Text analysis in the social sciences involves analyzing transcribed interviews, newspapers, historical and legal documents, and online data.
- Major approaches to text analysis include analysis of discourse positions, conversation analysis, CDA, content analysis, intertextual analysis, and analysis of texts as social information.
- Advantages of Internet-based data and social science research methods include their low cost, unobtrusiveness, and use of unprompted data from research participants.
- Risks and limitations of Internet-based data and research methods include limited researcher control, possible sample bias, and the risk of researcher passivity in data collection.

Review Questions

- What are the differences between text mining and text analysis methodologies?
- What are the main research processes involved in text mining?
- How is analysis of discourse positions different from conversation analysis?
- What kinds of software can be used for analysis of discourse positions and conversation analysis?

Discussion Questions

- If you were interested in conducting a CDA of a contemporary discourse, what discourse would you study? Where would you find data for your analysis?
- How do researchers choose between collecting data from offline sources, such as in-person interviews, and online sources, such as social media platforms?
- What are the most critical problems with using data from online sources?
- If you already have an idea for a research project, what are likely to be the most critical advantages and disadvantages of using online data for your project?
- What are some ways text mining research be used to benefit science and society?

Developing a Research Proposal

Select a social issue that interests you. How might you analyze how people talk about this issue? Are there differences between people from different communities and backgrounds in terms of how they think about this issue? Where (e.g., offline, online) do people talk about this issue, and how could you collect data from them?

Further Reading

Ayers, E. L. (1999). *The pasts and futures of digital history*. Retrieved June 17, 2015, from http://www.vcdh.virginia.edu/PastsFutures.html

Bauer, M. W., Bicquelet, A., & Suerdem, A. K. (Eds.), *Textual analysis. SAGE benchmarks in social research methods* (Vol. 1). Thousand Oaks, CA: Sage.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Kuckartz, U. (2014). *Qualitative text analysis: A guide to methods, practice, and using software.* Thousand Oaks, CA: Sage.

Roberts, C. W. (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts.* Mahwah, NJ: Lawrence Erlbaum.