Chapter 11

Amplifying Participant Voices Through Text Mining

Jonathan S. Lewis

https://orcid.org/0000-0002-5996-7131
uAspire, USA

ABSTRACT

Text mining presents an efficient, scalable method to separate signals and noise in large-scale text data, and therefore to effectively analyze open-ended survey responses as well as the tremendous amount of text that students, faculty, and staff produce through their interactions online. Traditional qualitative methods are impractical when working with these data, and text mining methods are consonant with current literature on thematic analysis. This chapter provides a tutorial for researchers new to this method, including a lengthy discussion of preprocessing tasks and knowledge extraction from both supervised and unsupervised activities, potential data sources, and the range of software (both proprietary and open-source) available to them. Examples are provided throughout the paper of text mining at work in two studies involving data collected from college students. Limitations of this method and implications for future research and policy are discussed.

INTRODUCTION

With the rise of internet-based communications, social media, and text messaging, researchers have access to new types of text data—sometimes produced intentionally, while at other times created as a by-product of social interactions—that can illuminate thoughts, behaviors, and emotions of students, faculty, and other key stakeholders in higher education. As the quantity of text data expands amid a “big data revolution,” traditional qualitative methods such as manual coding become impractical, and therefore researchers must find new methods for capturing participant voices and perspectives (Ignatow & Mihalcea, 2017, p. 5). Text mining, a form of data mining with roots in library science, information extraction, and natural language processing, is an ideal method to investigate new troves of unstructured data (Delen & Crossland, 2008; Ignatow & Mihalcea, 2017; Miner et al., 2012; Zilvinskis & Michalski, 2016).

DOI: 10.4018/978-1-7998-1173-2.ch011
This chapter is structured with minimal discussion of theory and maximal discussion of methods. The first section will propose briefly the efficacy of these methods and ground the argument in current literature, while the second section offers a comprehensive guide for researchers new to data mining. The goal is to equip the reader with the tools needed to begin investigating and analyzing text data. Examples will be provided throughout using anonymized data from two recent studies.

One study (Lewis, 2019), referred to hereafter as the leadership study, analyzed 67,790 responses to an open-ended prompt from the Multi-Institutional Study of Leadership (MSL). The MSL (2019) is a cross-sectional survey grounded in the social change model of leadership development as a theoretical framework. The survey collects a range of self-reported input, environment, and outcome data, in domains that include capacity and self-efficacy for leadership, as well as experiences before and during college that might relate to leadership (Komives, Dugan, & Segar, 2006). More than 350 institutions and over 610,000 students have participated in the MSL across seven cycles of the survey. The specific prompt that was analyzed for the leadership study requested that participants offer a definition of leadership in their own words. Across more than 400 variables, scales, and composite measures, it is the only open-ended query presented to all participants. One previous study (Haber, 2012) has examined participant responses to this question, using conventional methods of content analysis.

The second study, referred to hereafter as the college access study, offered an intervention of virtual college advising to high school seniors (CollegePoint, 2017). The purpose of this intervention was to use technology to help high-achieving, low- and moderate-income students apply to and, ideally, enroll at selective institutions. Across the high school classes of 2015 and 2016 combined, more than 9,000 students volunteered to receive college advising via text, email, phone, and video chat throughout their senior year. Twice during the intervention, participants were contacted by the National Opinion Research Center to complete a survey, which aimed to measure outcomes and opinions related to the intervention. Text mining was employed to analyze 4,691 responses to open-ended questions from the second wave of the survey, including one that asked participants to describe reasons they chose to enroll at a particular institution, and another that solicited suggestions for virtual advising.

BACKGROUND

Text mining is a form of data mining that can “turn text into numbers,” thus facilitating the efficient processing of large amounts of text data where traditional qualitative methods are impractical or inefficient (Miner et al., 2012, p. 30). In an educational context, text mining can be used to investigate a range of qualitative data provided by students through the teaching, research, and administrative functions of an institution (Zilvinskis & Michalski, 2016).

From application essays to course evaluations to survey responses, students leave a trail of text across higher education. As the academic environment expands to include new online spaces, students are providing increasing amounts of text data through everyday interactions with peers and instructors at their institution, across platforms as varied as course management systems (e.g., Blackboard, Canvas), blogs, e-portfolios, social media discussion boards, wall posts, photo captions, and tweets (He, 2013; Hung & Zhang, 2012; Rowan-Kenyon & Martínez Alemán, 2016; Zilvinskis & Michalski, 2016). Although not all text data is of substantive interest, researchers are beginning to examine a wide range of text thematically and generating implications for teaching, research, or practice. For instance, recent studies have explored reasons that students provide for withdrawing from a course (Michalski, 2014);