RE-THINKING STUDENT WRITTEN COMMENTS IN COURSE EVALUATIONS: TEXT MINING UNSTRUCTURED DATA FOR PROGRAM AND INSTITUTIONAL ASSESSMENT

A Dissertation Presented to the Faculty
of
California State University, Stanislaus

In Partial Fulfillment
Of the Requirements for the Degree
of Doctor of Education

by
Donald W. Jordan
May 2011

CERTIFICATION OF APPROVAL

RE-THINKING STUDENT WRITTEN COMMENTS IN COURSE EVALUATIONS: TEXT MINING UNSTRUCTURED DATA FOR PROGRAM AND INSTITUTIONAL ASSESSMENT

by
Donald W. Jordan

Signed Certification of Approval Page is

On File with the University Library

Dr. Oddmund Myhre Professor of Education	Date
Dr. Dennis Sayers Professor of Teacher Education	Date
Dr. Jace Hargis Assistant Provost	Date

TABLE OF CONTENTS

PAGE

PAGE

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

A nearly ubiquitous instrument of assessment for instructors and courses at the university and community college is the student course evaluation. One common feature of course evaluations is the open-ended questions that are often used to provide feedback to instructors on course and instructional content. Because of the difficulty in large scale assessment of written text, the written comments are often not analyzed with a systematic or consistent methodology. Technological advances, however, have made it possible to quantitatively study the unstructured data from these written responses through the algorithmic use of text and data mining. This study, using 835 surveys from a continuing education program over a 5-year period, employed an embedded correlational model using text mining methods such as Principle Component Analysis (PCA) and Singular Value Decomposition (SVD) within a qualitative framework to determine the viability of such an analysis on an institutional level. The study's major findings show that while there is only a weak correlation between the Likert responses and the open-ended written portion, there are significant words and patterns within the unstructured data that provide additional information at the institutional level. The results of this research suggest a need to rethink the design, implementation, and approach to the student course survey that can take advantage of text mining as an analytical tool for the institution.

CHAPTER I

INTRODUCTION TO THE STUDY

Introduction

Most colleges and universities conduct the ritual of course evaluations near the end of every semester. The evaluations are a common method of institutions to provide feedback to instructors regarding their classes. Course evaluations are a nearly ubiquitous instrument that is required by most colleges and have a longstanding role in education going back to 1918 when it first appeared in the literature (Coffman, 1954; Kilpatrick, 1918).

Alhija and Fresko (2009) state that student course evaluations have evolved into three basic forms, 1) a variety of statistical questions using multiple choice and Likert scale responses, 2) open-ended questions that allow students to respond with their own words, and 3) a combination of both. Most course evaluations are conducted using a simple survey form that is most often a combination of Likert scale responses to statistically valid questions or statements and one or more open-ended questions. The first section consists of anywhere between five and twenty questions using a Likert scale that asks students to rate various aspects of the course or instructor. Often, the answers correspond with a rating system that includes *Never, Almost Never, Sometimes, Almost Always, Always*. Alternately, it may have a series of questions that makes a statement and asks if the student strongly disagrees, disagrees,

neither agrees nor disagrees, agrees, or strongly agrees. Occasionally, the questions may be true or false, or multiple choice. The second part of the survey, then, consists of one to five questions (most often three questions) that are more general in nature and ask the student questions such as: "Please discuss the most positive/negative aspect of this course," and "What do you think went well in this course," or "What is one way that this course could be improved?" (Abrami, d'Apollonia, & Rosenfield, 2007; Sheehan & DuPrey, 1999; Abbott, Wulff, Nyquist, Ropp, & Hess, 1990).

Alhija and Fresko (2009) identify two basic purposes for these evaluations. First is to help faculty identify areas of both strength and weakness for the purpose of instructional improvement. Second is to help facilitate institutional decisions regarding tenure, course assignments, and faculty advancement. Because of the pervasiveness of use in higher education, there has been extensive research into the validity and reliability of such instruments (Renaud & Murray, 2005).

The statistical section of the survey is important for a number of reasons. First, it is easy to complete and requires very little effort on the part of the student. This is important as the shorter and easier a survey is, the higher the number of surveys are completed. Second, it ensures that all students are asked the exact same question, which then makes it easier to compare student's responses with each other, both within the class as well as between classes, instructors, and possibly programs. Such ratings can then be subjected to a variety of statistical calculations (the most common of which are averages and standard deviations) including chi squares and t-

tests to determine a variety of information about how students perceived their experience in the course or view of the instructor (Renaud & Murray, 2005).

Yet, as Hodges and Stanton (2007) point out, it is also a source of significant anxiety for faculty members. "Student ratings measure directly one product of instruction; namely, student satisfaction with teaching" (Abrami et al., 2007, p. 393). Student ratings seem to provide an indirect inference that higher rated instructors positively affect institutional offerings. Yet, this also raises fears among faculty "that one of the pernicious effects of student ratings is to make courses easy to court popularity" (Coleman & McKeachie, 1981, pp. 225).

There are limitations, however, to the statistical portion of the survey instrument. While it makes comparisons easier, and provides excellent quantitative data, there is a limit to the detail that can be teased from the data. Individual instructors can, for instance, place the numbers generated within a context, but once instructors are combined, the data becomes more general and less helpful for institutional level evaluations apart from ranking.

Another limitation of course surveys is that they are both informed and limited by the questions that are asked. Often, these questions are carefully vetted by administration and faculty groups or unions that seek to influence the nature of what can and cannot be asked. In some cases, student evaluations are designed and implemented by university Institutional Research (IR) departments, but often, they are created in conjunction with unions or faculty groups interested in protecting their members' (not always the students') interests, or more informally, by particular

departments or programs. It is not at all unusual for a university to have no one single standard instrument, but a variety created by different constituencies within the various schools, programs and departments. As such, often the questions that are asked may be indirect or at worst, superficial.

While there is much research on the use of multiple choice and Likert scale responses for course evaluations (Sheehan & DuPrey, 1999), little investigation has focused on students' written comments. The primary purpose of student comments is for individual feedback to instructors or for use in one-on-one evaluations with administrators for advancement, assessment, or tenure. Faculty that are interested in improving their courses often conduct their own informal surveys, either formative or substantive, but as Braskamp, Ory, & Pieper (1981) and Ory (2000) show there tends to be suspicion on the part of faculty as to the motives behind the administration's efforts to use student surveys for tenure and promotion.

Written Comments

Alhija and Fresko (2009) indicate that while there is research literature on student ratings and its validity and reliability in instructional and course improvement, there is far less research on the data obtained from the written comments and their relationship to the quantitative measures, despite the fact that they are included in most of the surveys that colleges and universities use. First, they are a catch-all for students to write out their observations, recommendations, frustrations, and any other issues that may not have been addressed. Some instructors

indicate that this section of the survey is the most important part as it gives a clearer picture of what the students really feel or think. Yet, as noted above, others are skeptical and have pointed out that students are not trained observers, have little knowledge of formal evaluation of teaching, and may not be privy to the pedagogy that the instructor is employing (Braskamp et al., 1981).

Student comments are problematic due to the fact that the literature suggests that as few as 10% of respondents provide written comments while other studies report a completion rate of up to 60%. There is some evidence that the method of evaluation may have an impact. Completion of the written open-ended questions is lower when using pencil-and-paper, while completion rates are much higher, generally by a factor of 2 or 3, when the evaluation is done on the computer or online (Aleamoni & Hexner, 1980; Aleamoni, 1987; Abrami et al., 2007). Writing out responses on the computer is seen as less of a barrier than writing out long-hand responses. Also, because it is open-ended, the text that is entered can range from a few noncritical words such as "The instructor is cool" to paragraphs of detailed analysis.

While the carefully constructed Likert scale questions on the first part of the survey are valuable for the statistical information that it reveals, the open-ended questions have always posed a problem for providing useful information at a program or institutional level. While the general form of the modern student survey has not changed much since it was first introduced in the early twentieth century, institutions have always been limited by what they could do with the information. Primarily, the

student's written comments were used to "fill in the gaps" or "illuminate" the meaning of the statistical data generated by the first part. Also, generally, the statistical data does little to answer the question of "why" for the instructor. The written comments, however, may be able to provide insight as to how a course was conducted, what went well, and what could be improved. But beyond the context of the individual course or instructor, little is done with such comments. Using traditional qualitative tools, it is difficult to imagine a large scale and sustained analysis of open-ended comments. What value could such an analysis provide?

Sheehan and DuPrey (1999) suggest looking at the relationship between the quantitative data and the students' written responses. Studies that have looked into such relationships suggest that written comments are generally aligned and correlate to the quantitative data. But such examinations are still relatively rare and tend to focus on comment frequency, length, content coding, and psychological characteristics of students who write them. The authors, in a two-year study of 3,632 course evaluations covering 161 psychology courses, found that the frequency of positive comments tends to be twice that of negative comments. Despite this, some of the main sources of anxiety for faculty are both comments that are unjustified, not critically constructive, or cruel, added to the fear that student responses could be used by administrators for political purposes or to make decisions based on hearsay that may not fully illuminate the entire situation.

Hodges and Stanton (2007) however, suggest that the written statements in student evaluations may reveal additional intellectual challenges common to novice

learners and may provide insight that is the cornerstone of scholarly teaching. Student written comments are generally more helpful in giving actual feedback to instructors than the carefully crafted statistical questions. There are limitations: low return rates, short unhelpful comments, irrelevant statements, off topic remarks. But the open-ended nature of the questions allows students to focus on what they felt was important. Often, this is the section that faculty look at first and it generally holds more meaning than the statistical information.

Prior to the 1990s, very little had been done to look at student comments in a systematic way using statistical methods (Abbott et al., 1990). Romero and Ventura (2007) show that because of increased computer processing power and availability, there is an increasing amount of research into using text mining (also called data mining) in educational applications. Text mining is an automated process that identifies relationships between text in unstructured data in order to reveal patterns, frequency, and predictive probabilities of subsequent relationships to other words and statistical data. The authors write that, "although the educational data mining is a very recent research area there is an important number of contributions published in journals, international congress specific workshops and some ongoing books that show it is one new promising area" (Romero and Ventura, 2007, p. 144). Chen and Chen (2009) furthermore suggest that the use of text mining might not only be used as a summative assessment of a class or instructor, but could also make it a practical tool in formative assessment to allow an instructor to assess the effectiveness of a class in the middle of the course.

Such open-ended comments by students have an advantage over statistical instruments in the fact that through the student's unstructured responses, issues and factors that may be significant may go unnoticed. But the fact that the data are unstructured poses problems that a traditional factor analysis and other statistical methods are not equipped to handle (Pan et al., 2009). Without the use of text mining tools or educational analytics (which Campbell, DeBlois and Oblinger (2007) define as a marriage of large unstructured data sets, statistical techniques and predictive modeling with the goal of producing actionable intelligence) a persistent and valid method of analyzing a large quantity of unstructured data generated by student course evaluation process in a systematic and consistent way, is lacking.

Problem Statement

An increasing number of studies over the last 20 years have taken a serious look at student written comments (Romero & Ventura, 2007; Abrami et al., 2007; Feldman, 2007). Student written comments have not received the same attention as quantitative data due to technical limitations on doing a reliable systematic study. The first studies to look at the written comments used traditional qualitative methods of coding a limited number of students or classes (Lin, McKeachie, & Tucker, 1984). In the 1990s, the advances of both computational power and the development of more sophisticated text mining techniques have allowed for a far more sophisticated analysis to be performed on a greater number of samples (Romero & Ventura, 2007; Abrami et al., 2007).

Unstructured data is difficult to process and analyze except in a limited ethnographic manner. The model of analysis for this study has two aims: 1) to show whether or not there is a quantifiable correlation between the Likert scale responses of a course evaluation and the results of the PCA analysis and sentiment analysis of the students' written comments, and 2) whether or not the expressed satisfaction or dissatisfaction in the way they formulate responses provides actionable information to the institution that is not identifiable from the Likert scale responses. If there is a significant correlation in both areas, this could have a significant impact on the way that course evaluations are designed. Current instruments are limited in their design to look for presence and particular information. Through text mining, open-ended questions might bring to light aspects of evaluation that currently cannot be explored.

Methodology

Using a modified adaption of a text mining workflow (Luan, Zhao, & Hayek, 2009; Nisbet, Elder, & Miner, 2009) the test will be processed using PCA analysis in order to create a SVD for *k* best predictors in order to create an additive index for comparison with the Likert scale responses. The text will be coded algorithmically into positive and negative connotations and the results compared to the results of a hand-coded qualitative process using an embedded correlational model (Flick, 2006).

Nature of the Study

The primary purpose of this research is to first determine a correlation between the quantitative portion of a set of student course evaluations and the PCA

analysis of the written comments on those evaluations (Braskamp et al., 1981; Alhija & Fresko, 2009; Nisbet, Elder, & Miner, 2009). This can be summarized in the following research questions:

Research Question One: Are the student comments of course evaluations aligned to the quantitative portion of the course evaluation instrument?

Hypothesis One: The student comments of course evaluations are aligned to the quantitative portion of the course evaluation instrument.

Research Question Two: Are there words and patterns prevalent in the unstructured data of student comments of course evaluations that can classify individual courses on the basis of negative connotations?

Hypothesis Two: There are words and patterns prevalent in the unstructured data of student comments of course evaluations that can classify individual courses on the basis of negative connotations.

Research Question Three: Are there words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can provide additional information at a program or institutional level?

Hypothesis Three: There are words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can provide additional information at a program or institutional level.

Research Question Four: Is there an association between the results of the text mining analysis of the unstructured data and a qualitative analysis of the unstructured data?

Hypothesis Four: There is an association between the results of the text mining analysis of the unstructured data and a qualitative analysis of the unstructured data.

Operational Definitions

Algorithm. A set of automated instructions, operations, or procedures that produce a similar outcome.

Inverse Document Frequency (IDF). The process of transforming a raw frequency count of words in a collection of documents to simultaneously express both the frequency as well as the extent to which a particular word is used in only specific documents in a collection. (Nisbet, Elder, & Miner, 2009).

K-Means algorithm. An algorithm that is used to assign K clusters as a representation of N points resulting in $K<N$. The points are iteratively adjusted so that every N point is assigned to one of the K clusters where each of the K clusters is the mean of its assigned points (Bishop, 1995).

Knowledge discovery. The process of using an algorithm to automatically search large volumes of unstructured data for patterns that can be described as "knowledge" about the dataset (Nisbet, Elder, & Miner, 2009).

Principle Component Analysis (PCA). PCA analysis is a technique used to identify some strong predictor in a data set. It is used for revealing relationships between variables by identifying a group or cluster of principal components in order to reduce the quantity of original variables in a data set (Nisbet, Elder, & Miner, 2009).

Singular Value Decomposition (SVD). An analytic tool that can be used to extract underlying dimensions or components that account for most of the common contents or meaning of the words that were extracted (Nisbet, Elder, & Miner, 2009).

SVD Word Coefficient. The value of a word in a list of frequencies that is assigned relative importance by both its frequency as well as its relationship to other words in the dataset.

Text mining. An algorithmic process that includes 1) the identification of sets of related words in a document or a series of documents, 2) the process of creating word clusters, 3) the identification of clusters with some variable or component, 4) exploratory analysis using both structured and unstructured data to perform knowledge discovery to find hidden patterns, and 5) identification of frequent item sets and word clusters.

Sentiment analysis. A process that attempts to algorithmically determine the attitude, judgment, evaluation, or emotional state of a writer or a speaker on a given topic (Liu, 2010).

Conceptual Framework

The evaluation of courses and instruction is a complex and important matter. Alhija and Fresko (2009) identify two purposes of evaluation activities as to inform instructional improvement and to assist in decision making activities such as advancement and tenure. Yet there is a complex relationship between the effectiveness of instruction and the perceptions of the student. The authors identify in their model depicting components of student written evaluations (Figure 1) an intricate web of relationships inherent in student observations.

The Alhija and Fresko (2009) model tries to explain the context within which the student evaluations take place. Often, evaluations ask students for specific information on one aspect of their experience, most often relating primarily to the instructor or the course and its design. The model in Figure 1 shows another important influence on the student when they are evaluating their experience: namely the context of the institution.

In the standard statistical portion of the evaluation, the focus of the questions can be effectively set to evaluate one or all of these major influences or primary objects. The instructor, as a construct, is made up of several subattributes such as personal traits, teaching style, classroom management, organization, and other attributes under the general evaluation category. Courses, similarly, can be judged by a variety of subattributes such as content, assignments, and objectives, among others. Alhija and Fresko (2009) add a third category called context, to which they attribute scheduling issues and student composition, but we could just as easily add department

or program issues and attributes. The important thing to note is that the evaluation of the instructor is not divorced, in the general attitudes and feelings of the student, from

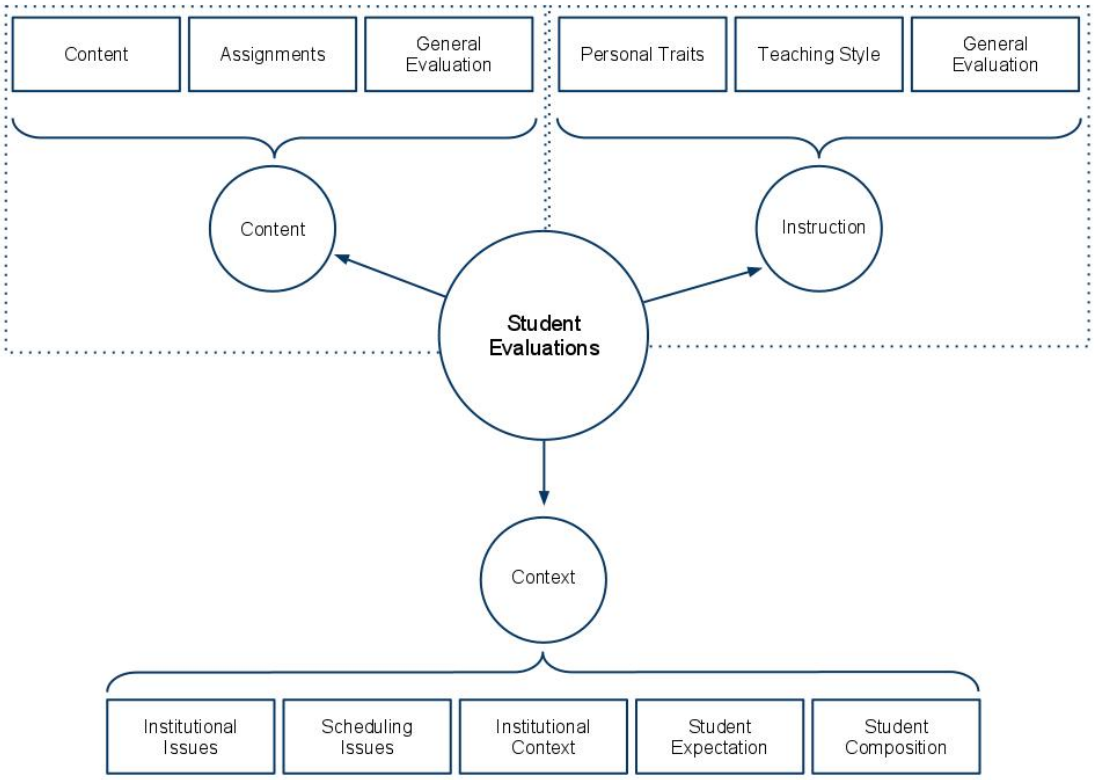

Figure 1. Model depicting components of student written evaluations (Alhija & Fresko, 2009).

the course, the program, the department, or the institution. Programming and scheduling issues may both positively and negatively impact the evaluation or perception of the instructor or course. Furthermore, the uses of this model become important in both PCA analysis and sentiment analysis when evaluating unstructured text because all of the influences that affect a student's evaluation must be identified prior to the analysis so that the unstructured text can be arranged and categorized in a structured and meaningful way.

Assumptions, Scope, and Delimitations

According to the literature, accuracy is still an issue with sentiment analysis and suffers from some problems (Liu, 2010). First, there is the question of precision or how accurate the discovered opinions are. Next, is recall, or how much is left undiscovered. For example, which statement is better? 1) The instructor is great, or 2) This was my second history class and I felt that the instructor really made the information accessible and understandable. While the second example appears to be more informative than the first, it is still lacking critical information such as how the information was made accessible. Was it the instructor, her organization, or the course design? Despite this, the primary advantage in text mining is that it is not the specifics of individual statements that are examined, but similar relationships and statements that are reflected longitudinally.

Significance of the Study

Students at the institution have contact with many dozens of faculty and staff. These students are enrolled in a variety of courses and have their own unique perspectives, experiences, and goals. No one student may be in the position to evaluate an entire course or program, but the words that they use, the sentiments that they write about, do carry a lot of meaning that may be useful in aggregate. Systems such as these can also have the benefit of combining more than just traditional course evaluations. If there were a system that allowed for students to post reviews and comments at the college website, and algorithmic monitoring of course evaluations

along with other micro and macro blogging such as Twitter, universities could have a real-time window on the pulse of the university with data that could track trends and alert officials and instructors to problems in a type of early warning system. Student written comments are often thought of as primarily useful only as they relate to a specific instructor or class. With text mining or perhaps even sentiment analysis, useful, actionable information may be uncovered that may be used in institutional assessment, program review, and faculty and staff development.

Implications and Challenges for the Future

Some of the challenges that are present in using text mining to improve programs, departments, and institutions revolve around a number of factors that have been present for at least the last century. First, there is the actual data collection, how, when, and where the data is collected. Then there is the issue of how to encourage students to complete the survey responses. Next, how will this data be used? There can be some distrust in many organizations between faculty and administration, therefore, while it certainly adds additional complexity to the situation, institutions should be careful and develop a comprehensive policy regarding the use of such data. Colleges and universities, over the last few decades, have become very good at generating data as accrediting bodies and governmental agencies have become increasingly vigilant and more refined in their efforts to make sure that higher education institutions do what they say they are doing and increasingly require them to be accountable.

Many studies have shown that there is an effect on the data regarding when and how the data is collected. Students participate more when they feel the instructor has time to address their issues, so response rates are higher if done in the middle of the semester. However, from an institutional perspective, such data is not as valuable because the students have not yet experienced the full class at that point, therefore, favoring the end of semester evaluation (Albbot, Wuff, Nyquist, Ropp, & Hess, 1990; Abrami et al., 2007). The length of the survey is also a factor—generally shorter surveys produced better response rates. Additionally, if students are instructed that the survey is used for evaluation and tenure review purposes, they tend to be more responsive and positive while if they are told that it is just for the instructor and that the feedback is used to improve a particular course, then they tend to be less positive and more critical. An additional complication is the fact that there are two very distinct purposes in student evaluations: the evaluation of the instructor and the evaluation of the institution, and both of these evaluations are based on opinion (Alhija & Fresko, 2009).

But opinions may be important data points for making decisions, and the more opinions that are registered, the better text mining is at revealing the underlying structure of the data. In the past, unstructured text that students might have written on course evaluation forms was accessible only to the few interested individuals who took the time to read them. But there is no way to aggregate all of the opinions together. With text mining and sentiment analysis, global scale opinions are no longer limited to small-scale surveys, focus groups, or researchers.

These conflicts will continue to remain as long as the current form of evaluation continues—one based on scarcity and limitation. The statistical nature of the student survey as it is now employed operates on the assumption that specific agents are being evaluated (which can be optimally targeted though the use of well-crafted questions, placement of the survey within the semester, and a framework of reasoning given to students when filling out the survey, among other things). Text mining has the potential to change many of these assumptions. While statistical approaches attempt to glean information out of a little evidence, text mining does, essentially, the opposite. It attempts to use a little information out of a lot of data. It is true that the responses in student surveys are often the least frequently filled out and often are only completed less than 20% of the time (Aleamoni & Hexner, 1980; Aleamoni, 1987; Abrami et al., 2007). Yet, in this model, students are often not prepared to fill out these surveys and are given only one chance, usually at the end of class when they want to go home. Would students respond better just after an event that they felt was important, whether it was good or bad? One of the freedoms of text mining is that we could expand our notion of the student survey for an event that happens at one point in the semester to something that is continual and always ongoing. This would, however, require a drastic rethinking and restructuring of the student feedback instrument.

In statistical approaches, the questions and framing of those questions are important as the observer is trying to infer meaning from a limited dataset. Text mining suffers from no such limitation. In fact, the questions become largely

irrelevant because the miners are looking for patterns. There needs to be no specific question at all, the only thing that is important is that there is an abundance of text (Nisbet, Elder, & Miner, 2009). Perhaps the whole notion of course surveys becomes irrelevant in the age of text mining. For example, if there were a discussion board on all class online shells that was labeled: questions, problems, and successes, and students were encouraged to go there with all questions and problems, this would give the institution much of the same data that would be given in the end-of-class evaluation with one major difference—it would be constant and ongoing. For privacy concerns, such comments could be filtered in such a way as to limit its use at the program and institutional level. Issues within a single course could be confined to those the institution deems necessary such as the instructor and possibly the dean, but issues that trend across programs, departments, and the entire institution would be illuminated in real time and would allow better decisions to be made more quickly.

As another example, suppose there is high student frustration with the current Learning Management System (LMS) and students are having a difficult time finding their coursework, or assignments are disappearing, and it is a generally frustrating experience. This might be an issue that could be raised on the course evaluation form; however, because the statistical questions do not specially ask about a class's online component, then it may only be reflected in more dissatisfaction on the instructor's ability to organize the class. A few students might mention this on the written portion, but this will largely only be read by the instructor, who may either write it off later as an information technology (IT) issue, or internalize it and feel that it is his or her fault

and may or may not seek additional training—assuming that it is something that could be helped with extra training or that the instructor is sufficiently motivated. Text mining can change this issue. If the evaluations were consistent and ongoing, such issues would be elevated because 1) many students are indicating similar notions about the LMS or, 2) perhaps 20 instances across 10 classes involved assignments disappearing (perhaps a training issue, but then again, perhaps a configuration issue on the part of IT), and there is a common complaint that the navigation is difficult that appears universally across the campus. With such data, the administration would be in a better position to reassure both faculty members and students, and make the determination as to whether this was an issue for increased training, or if a new system needs to be developed or purchased. Other global issues regarding classroom management as well as teaching methods could highlight faculty development needs as well as students' desires for what might be missing.

The above example shows an important aspect of using text mining in instructional evaluations. First, because a wide net is cast, and there is no pre-determined type of responses required or specific information asked for, a variety of issues can come to the surface. Often, colleges and businesses are blindsided not because they were not paying attention or ignoring information, but because they were not asking the right questions. In the statistical approach to evaluations, the researcher needs to ask the right questions, or the data is irrelevant. With text mining, one is able to ask more general questions such as "How do you feel?" and ask them more often and in a more persistent manner. One additional highlight from the

example above is that such evaluation methods no longer only focus on a particular class or instructor. The entire institution is under review, including the students. Additional predictive analysis (a particular strength of principle component analysis) could be done to determine if certain issues trend higher in students who are leaving or fail to come back, or if there are spikes in drops just as there is an increase in certain types of complaints. It can also illustrate issues that are going well and areas where the college is being particularly effective.

Ideally, such a system could be automated. This may not be feasible in the near future, but with training, commercially available and open source systems could provide a wealth of information to educational institutions that is currently inaccessible on a large and systematic scale.

Summary

Student surveys are a nearly universal tool in higher education. While there is no one standard survey, almost all such surveys contain questions that allow for simple responses on a Likert scale. Additionally, many institutions use forms of the instrument that contain open-ended questions that are of little use to program and institutional review due to the fact that 1) historically, there is a very low completion rate of open-ended questions, and 2) it is difficult to quantify such a large number of unstructured responses in a way that is consistent or useful on a program or institutional level. However, text mining may provide a method of processing a large amount of unstructured text, such as open-ended student comments, in order to

provide institutions with relevant (and hopefully actionable) information that is useful to not just the instructor, but to the program and institution as well.

The next chapter will provide background for this study by examining some of the issues surrounding student course surveys that will frame the literature on the validity and reliability of the instruments. The chapter will also examine how purpose and structure of the instruments affect both the results and the perception of those results. Text mining and sentiment analysis will be discussed as a strategy for providing additional contextual information to inform or frame the statistical data.

Chapter III will be concerned with the methodology of this research with an overview of the correlational model used to assess the unstructured data presented in the student written comments using both text mining and qualitative techniques.

Chapter IV will present the description of the data set used by this study along with the details of the text mining setup followed by the analysis of findings for each of the research questions.

Finally, in Chapter V, the results will be interpreted for each of the research questions and the results related back to the literature along with recommendations and suggestions for future research.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

The student evaluation, practiced in some form by most universities and colleges is a very well documented and studied tool, however, its origin is relatively recent, going back to 1918. Though repeated studies have upheld the validity and reliability of such interments, the current environment is adding emphasis on good teaching even as teacher-centered methods are falling out of fashion and student-centered learning is becoming more prevalent (Pan et al., 2009).

Because of this shift in focus, Pan et al. (2009) argue that student ratings will continue to be important in not only personnel decisions and summative evaluations, but also in formative assessment and, perhaps tangentially, in faculty development efforts. As more emphasis is put onto student feedback, there is a limit to the number and type of questions that can be effectively included in a quantitative instrument as well as a limit to what can be reasonably demanded from such instruments. While basic statistical analysis provides quite a bit of useful information, statistical methods that result in a quantifiable number are 1) only as informative as the instrument design allows, and 2) open to potential abuse and personal interpretations that could undermine the validity of the process.

Text mining techniques such as PCA and sentiment analysis applied to the written comments from open-ended comments may provide an alternate method of extracting detailed and complex information from largely unstructured data sets (Pan et al., 2009).

This chapter will begin by looking at some of the prominent myths, half-truths, and fears of student course surveys. Next, literature surrounding validity and reliability is discussed along with the purpose of course surveys and the common types of evaluation instruments. The effects of timing and methods of collection that have been shown to affect results will also be reviewed. The chapter will conclude with an overview of the methodology employed to create actionable information from unstructured data using text mining along with an overview of PCA and sentiment analysis.

Review of Research and Literature

Introduced in the literature in 1918 (Kilpatrick, 1918), and more commonly adopted and studied in the 1930s and 1940s (Coffman, 1954; Guthrie, 1954) and often conducted voluntarily, student evaluations have become one of the most common ways for schools to assess instructors and courses and are often required at most institutions. While Coffman (1954) does not specifically deal with the written comments within course evaluations, he acknowledges the emotional aspect of the issue of having students rate their instructors. Despite the evidence he presents regarding both validity and reliability, he argues that this issue is complicated by a

resistance on both sides towards evidence both critical of supportive to the idea of students evaluating their instructors. However, he states "the picture one gets of students from this analysis is not one of incompetent judges assessing teaching on the basis of superficial evidence" (Coffman, 1954, pp. 284). While acknowledging the problematic nature of seeking student opinion, he argues that the results of such surveys provide valuable information on what students value and what they are looking for. Any deficiencies with course survey instruments, he argues, are more the result of inadequate or ill-constructed questions than the student's ability to make observations. The author implies that such information might help inform improvement in instruction; however, it stops short of suggesting any sort of faculty development or administrative development regarding these issues.

Myths, half-truths, and fears. Feldman (2007) and Pan, et. al. (2009) address many of the myths, half-truths, and fears surrounding the subject of student evaluations that are important caveats, but will not be directly addressed in this study. Feldman refers to these myths, half-truths, and fears as one or more factors believed to inappropriately influence the judgment of students about their professors or courses. But the real question, he argues, is whether a condition or influence actually affects professors and their instruction (nonbias) or if an influence only affects the students' perceptions of instructors and their courses in such a way that they do not accurately reflect their actual experience.

The authors have found that the following are untrue or misleading because there is little, if any, supporting research (Abrami et al., 2007; Feldman, 1976, 1989, 2007; Marsh, 1984, 1987, 1989; Pan et. al., 2009):

- Students are not capable of making judgments about instruction because of a lack of maturity or experience. In fact, Braskamp et al. (1979) show that there is a correlation between student ratings, instructor's ratings, and course achievement.
- Only colleagues with expertise (and publication records) are qualified to evaluate instruction.
- Research and teaching skills are closely aligned to the point that it is unnecessary to separate the two in evaluations.
- Student ratings and evaluations are nothing more than popularity contests with the popular instructor having the best personality and giving the highest grades. Aleamoni and Hexner (1980) found in a review of the literature that in the period between 1934 and 1970, 23 studies showed no relationship between grades and expected grades and positive ratings in student evaluations. In the same period, 27 studies did find a positive relationship, however, the effect was weak with a median correlation of 0.14, a mean of 0.18 and a standard deviation of 0.16.
- Students are unable to judge a course or instructor honestly until after they have been away from the course (and possibly the university) for years.

- Student ratings are unreliable and invalid. This is, in fact, refuted by most of the literature in the field. In cases where surveys were deemed invalid, the issues was more about poor design of the instrument than the fault of students or the nature of using the course survey for feedback (Feldman, 2007; Abrami et al., 2007).
- Factors outside of the instructor's control, such as time and day of the course, class size, level of difficulty, and expected grade affects ratings (Marsh, 1987; Feldman, 2007).
- Students cannot meaningfully be used to improve instruction. In fact Coffman (1954) and Gallagher (2000) both argue that the use of course survey data not only provides meaningful feedback, in aggregate over time, but that the instructor must be a willing participant in the process.
- Gender of a student and instructor affects ratings (though Feldman (1976a 1976b, 1989b, 1993) found that there was a slight advantage favoring women (average r = .02) as to be insignificant).

Some research has been done that does look into what contextual factors go into influencing student satisfaction measures (Downey, 2003; Manochehri & Young, 2006). Downey's study identifies that student satisfaction is influenced by gender, emotional awareness (Salovey, Mayer, Goldman, Turvey, & Palfai, 1995), and awareness of physiological states (Schachter & Singer, 1962; Reisenzein,1983). Lim et al. (2008) and Kember and Wong (2000) discuss the importance of the teaching

method on student satisfaction. One of the most often-cited overviews of research on student ratings of instruction is Marsh (1987) who states

> Research described in this article demonstrates that student ratings are clearly multidimensional, quite reliable, reasonably valid, relatively uncontaminated by many variables often seen as sources of potential bias, and are seen to be useful by students, faculty, and administrators. However, the same findings also demonstrate that student ratings may have some halo effect, have at least some unreliability, have only modest agreement with some criteria of effective teaching, are probably affected by some potential sources of bias and are viewed with some skepticism by faculty as a basis for personnel decisions. It should be noted that this level of uncertainty probably also exists in every area of applied psychology and for all personnel evaluation systems. Nevertheless, the reported results clearly demonstrate that a considerable amount of useful information can be obtained from student ratings; useful for feedback to faculty, useful for personnel decisions, useful to students in the selection of courses, and useful for the study of teaching. Probably, students' evaluations of teaching effectiveness are the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research (p. 369).

The information gleaned is important, varied, and open to interpretation. While many of concerns may be valid, there is still a wealth of important and actionable information that can be taken from a well designed course survey.

Validity. There is a significant amount of research on the validity and the reliability of the quantitative portions of the student evaluations (Ory, 2000). The research on using the qualitative (open-ended) portion of the evaluations as a valid or reliable tool is far more scarce, despite an acknowledgement that such open-ended questions provide a broader spectrum of information that could be used (Lewis, 2001). Aleamoni and Hexner (1980) addressed the fact that validity of student surveys is far more difficult than assessing reliability. For reliability, they suggest that the correlations between one year and the next range between 0.80 to 0.90, which is consistent with the literature. For validity, statistical tools such as factor analysis can be used to verify subjectively determined dimensions that the survey covers such as instruction, setting, and course.

Research conducted in the 1980s was primarily concerned with establishing the validity of student ratings and what such data really meant in practical and political terms. Braskamp et al, (1981) approached student written comments as an issue of validity. Their primary purpose was to document the validity of the written comments with the overall ratings as measured by the Likert scale questions on the instrument. Their approach was to randomly select 60 courses from a possible 2,685 courses with restrictions to ensure that each of the university's 10 academic departments were represented, that each selection had enrollments of 10 to 50 students, and that there was an even distribution of instructor rank and experience represented by the sample. Overall, a total of 3,240 comments from 924 students were logged and coded into 22 categories relating to instruction, grading, and the

course as well as an additional category of positive and negative. They found that the diversity of the written comments matched the range of categories covered in the Likert scale question, though they did not speak to comments that may have gone outside the range of their instruments. They ran a frequency analysis of positive and negative comments grouped by the three major categories and objective ratings which showed a high degree of convergent validity between the comments and ratings.

Centra (1987) looked at a variety of evaluations such as student ratings, colleague evaluation, and evaluation of research and scholarship, and was primarily concerned with the definition of good teaching. There is evidence of concern over whether effective teaching was subordinated by student evaluations to an instructor's personality, entertainment skills, or other factors not generally associated with good teaching. However, many studies confirm the validity of student ratings by checking the ratings against measures of student learning across multiple institutions and a wide range of teachers and instructional levels (Aleamoni & Hexner, 1980; Marsh, 1892, 1983, 1984, 1987, 1989; Gallagher, 2000).

Research in the 1990s began to show concern that student ratings data were either overused or used inappropriately (Franklin & Theall, 1989) but overall, these concerns did not invalidate earlier findings (Marsh, 1987; Theall & Franklin, 1990, 1999). Kember and Wong (2000) reaffirmed that such data could be used as a primary source for an evaluation of instruction, though they acknowledged that the results are biased by students' conceptions of learning and perceived purpose of education. In their study, they sought to see if course evaluation surveys really did

accurately reflect student's perceptions of teaching. Through the use of interviews, the researchers sought to gain a perspective on students' perceptions of both good and poor teaching. Semi-structured interviews were conducted with 55 students which focused on a variety of contextual areas: courses and teaching methods, workload, study habits, peer support, out of class learning, and classroom activities. Once coded, they were analyzed and compared to responses and ratings in the course surveys. The analysis showed that what students believe about teaching and learning fell upon a continuum from passive to active—preferences which they showed had a predictable effect on students' concept of teacher effectiveness. This explains, the authors argue, inconsistencies in the data where students disagree with one another over teaching effectiveness by rating an instructor high while another rates the same instructor lower on the scale. Student preference for viewing learning as passive or active helps account for such discrepancies. While the authors used interviews in their study, open-ended questions may also provide contextual clues for such preferences.

Kember and Wong's (2000) conclusions are consistent with Saljo's (1979) set of descriptors for students' concept of learning as one or more combinations of the following: 1) quantitative increase in knowledge, 2) memorizing, 3) acquisition of facts, 4) abstraction of meaning, and 5) an interpretive process aimed at understanding meaning. Marton, Dall'Alba, and Beaty (1993) later added to this list with a sixth factor: changing as a person. But overall, such ratings were shown to be a valid source of determining instructional quality to be used in making informed decisions about hiring, tenure, promotion, and funding. This then raises the question

regarding the importance of knowing what students mean by good and bad teaching when they rate instructors on a Likert scale and, when they make that judgment, to what extent are they influenced by their own perception of learning as a fundamentally active or passive activity.

Prior to the 1990s, very little research was done on the open-ended comments that students provided in most evaluations. Lin et al. (2008) discussed the fact that student written comments were not easily summarized due to the non-standard and idiosyncratic nature of the responses. Because of the difficulty summarizing such comments in a fast and systematic way, the authors suggested that such comments were useful only for returning to the instructor for their own analysis and consideration. By the mid-1990s, limited analysis was conducted using newly emerging tools such as text mining which allowed for systematic analysis of unstructured and open-ended textual responses (Romero & Ventura, 2007).

In the late 1990s and later, emphasis began to shift from teacher-centered approaches to evaluation as student-centered learning (Abrami et al., 2007). But with the proliferation of online systems, and the rise of for-profit online institutions, faculty have become increasingly concerned about possible misuse or misinterpretation of evaluation results (Feldman, 2007)

Purpose in evaluation of courses and instructors. In higher education, student feedback has become indispensable. Though it is rarely the only means of evaluation, its significance is undeniable. Gallagher (2000) shows that well-constructed instruments provide both valid and reliable information on teaching

effectiveness. Using Cooley's self-theory, Gallagher argues that for course surveys to be of most use to instructors, the instructor must take an active part in listening to and interpreting the evaluations in the context of the long history of research on student evaluations as well as act upon both positive and negative feedback in ways that result in demonstrable improvements in teaching. But, he acknowledges, this requires a culture and emphasis on teaching as a process of improvement rather than high stakes personnel assessment. Moreover, surveys, he argues, fit within a framework of other improvement strategies such as mentoring, research on teaching effectiveness, and large-sample analysis of student evaluations for an instructor's department.

Most evaluations provide quantitative data that can easily be graded or scaled. Many, though not all, instruments provide a qualitative section which solicits comments and observations from students. The rational for this is to encourage students to highlight other areas of strengths and weaknesses that the quantitative questions may have missed (Pan et al., 2009) but is also useful in providing an overall context.

Types and structure of evaluations. While there is a great range of student evaluation instruments currently being used, Sheehan and DuPrey (1999) found that they fall into three general categories. The most common instrument is a purely quantitative questionnaire made up primarily of Likert scale items. There are also instruments that have only open-ended questions, though many are made of a combination of the quantitative scales and open-ended questions. Abbott et al. (1990)

also discuss the use of student interviews as an additional, though less common, method of data collection.

Effects of timing and methods. A few factors make such surveys more or less useful. First, completion rates vary due to the length of the survey, but Abrami et al. (2007) also show that the timing and method of collecting an evaluation has an effect on the results. Most schools still rely on pencil-and-paper surveys that are handed out in class toward the end of the semester. Students often see this as less valuable because even if the instructor is interested in their comments, the perception is that it will have no effect on this semester or their own education. The lowest satisfaction occurred with evaluations at the end of the course with limited or no instructor responses to the students' feedback which suggests that students value the response of the instructors in their evaluation and feel that their comments are more useful when given in such a way as to have an effect on the current class. No response by the instructor to the evaluations results in the least amount of satisfaction for the student (Abbott et al., 1990; Abrami et al., 2007). The authors also found that students' initial perception of the purpose of the course evaluations and what they would be used for had an effect on the ratings. Students are generally more critical if it is thought only the instructor will be reading the evaluation (in the case of where instructors seek feedback to improve despite any institutional requirement to do so) but will be generally better if the students are informed that personnel and administrative decisions will be made based partly on the results of the survey.

In one study, 244 students were given courses surveys. The students were split into two random groups and analyzed to ensure that they did not differ markedly from each on a variety of measures. The first group was given evaluations and told that only the instructor would be reviewing the evaluation and would use the surveys for course improvement while another group was told that the information would be used for administrative purposes such as performance evaluations and would affect tenure, promotion, and salary considerations. Overall there was a 0.27 mean difference on a 5-point scale between the two groups with the greater satisfaction expressed on the evaluations that the students perceived as being more important by being used for administrative purposes (Aleamoni & Hexner, 1980; Aleamoni, 1987). The authors hypothesize that students may feel more candid when providing feedback to only the instructor, but may self censor minor criticism if there is the possibility that it could have negative repercussions outside the class. While there are a few interpretations of this data, it is probable that while students may not be shy about discussing the shortcomings of their instructors in these evaluations, many, if not most, do not wish to see the instructor harmed by their comments so are more willing to give a better review on paper, while possibly bringing up dissatisfaction either in person or a variety of other ways that may be seen as more constructive, or perhaps even remaining silent, feeling that silence is better than getting someone in trouble over an issue that may or may not be a big deal. Therefore, if the student perceives that the instructor really wants feedback, and that the feedback can genuinely improve the course, then there is more of an effort to be frank. If, on the other hand, it is seen

as evaluative on the part of the instructor, then they may be more inclined to play along, not wishing harm in a process that will have no effect on their own education. Overall, Abrami et al. (2007) conclude that students are more satisfied overall when the student surveys are conducted at midterm with an extended instructor response with a demonstrated effort toward quality and improvement.

Hardy (2003) conducted a review of research evaluating Northwestern University's move from pencil-and-paper surveys to online surveys. Initially, in 2001, a study of over 5,000 courses in the four quarters after the introduction of online surveys, showed that the ratings were 0.25 lower on a 6-point scale for online feedback, thus feeding into the perception that results would be lower for online collection. To further examine this issue, a study was conducted with seven sections of large introductory courses raging in size from 432 to 634 students with five courses being evaluated using pencil-and-paper and two sections being evaluated online. This time, ratings gathered online were shown to be consistent with those collected by paper. Furthermore, the research showed that students wrote substantially more comments when submitting online, though there was little difference in the type of responses—positive, negative, and mixed—between methods of delivery. Overall, the research found that while overall response rates were lower for online submission, students who submitted online surveys were almost twice as likely to provide written comments.

Johnson (2003) also examined whether the method of delivery for the course survey affected student responses. The author looked at the completion rates of

pencil-and-paper surveys verses surveys conducted on the computer. While initially lower than pencil-and-paper, the response rates for the computer surveys at Brigham Young University eventually surpassed the pencil-and-paper over a period of six years. While the authors hypothesize that this is due in part to increased access to computers, and greater familiarity with electronic tools over that period, the situation illustrates the role that external factors play on student responses.

There is also some evidence that, despite some practical and political issues, making student evaluations of courses available to other students can have a positive effect in guiding students toward courses in which they are more likely to be successful. A study by Coleman and McKeachie (1981) showed that when students are given access to the results of student evaluations, the most highly rated course is more often elected, in comparison with the control group, despite the course requiring more work. The authors state that "many professors have contended that one of the pernicious effects of student ratings is to make courses easy to court popularity" (Coleman & McKeachie, 1981, pp. 225). The feeling is easy to understand, despite evidence that there is no correlation between grade or expected grade and overall ratings.

Abbott et al. (1990) found that the method and timing of the course survey may have an effect on student satisfaction. In their study, they surveyed eight different procedures for collecting student satisfaction by varying three factors: method (interviews vs. paper), timing (midquarter vs. finals), and instructor reaction (if allowed) in a 2 x 2 x 2 design. Each procedure was conducted on similar classes

with an enrollment of greater than 20. They saw that timing had a large effect with student indication of satisfaction. For midterm collection with group interviews (M = 4.70) and the mean satisfaction for individual responses (M=4.39) were significantly greater than the end-of-quarter collection with group interviews (M=4.07) and standard individual responses (M=4.09). Similar differences were also found in the response to the question, "How satisfied are you with the instructor's response to the information provided by students about this course?" The authors argue that this shows a clear indication that collection of survey data at the midterm has a strong influence on student satisfaction at the end of the course across all eight methods of collection. Across all methods, the authors found that collection at midterm with extended instructor feedback (where the instructor would address any issues in class that was rated in the surveys) led to the greatest student satisfaction at the end of the course.

Methodology

Yu, DiGangi, Jannasch-Pennell, Lo, and Kaprolet (2007) see text mining as a key tool in the issue of retention and persistence. The authors use text mining techniques to determine if students who commute and take online classes are missing a key element of interaction that may help many traditional students stay enrolled.

Pan et al. (2009) used SPSS Text Analysis for Surveys (STAS) to provide analysis of unstructured data in a timely manner that would make ongoing analysis both possible and practical. They used written comments from two questions, "What

are the teacher's strengths?" and "What improvements would you suggest to the teacher?" (p. 79). From these two questions, the qualitative data can be used to strengthen the quantitative data with systematic and meaningful qualitative interpretation which, the authors argue, can provide widespread benefits and applications such as

a) Creating a profile of positive descriptors detailing what students regard as an effective teacher;
b) Creating a profile of negative descriptors detailing what students regard as an ineffective teacher;
c) Evaluating the strengths of an individual according to the (a) profile; evaluating the weaknesses of an individual according to the (b) profile;
d) Providing a personal profile (fingerprint) of strengths for an individual;
e) Providing a personal profile (fingerprint) of weaknesses for an individual;
f) Offering a basis for educating students about "effective teaching" in relation to "effective learning", based on (a) and (b);
g) Enabling the use of information available from (c), (d), (e), and (f) for summative purposes;
h) Enabling the use of information available from (b) for identifying possible corrective measures for the entire teaching community; providing information available from (f) to enable an individual to improve his/her teaching performance;

i) Using (a) and (b) as a baseline to enable the institution to study the developmental

j) Aspect of students' concepts of teacher and teaching over time; documenting (c), (d), (e), and (f) to provide detailed profiles of teaching staff in the institution and allow comparison over time (Pan et al. p. 98, 2009).

Additionally, the above model provides a suitable model for creating actionable and reportable information in a variety of contexts from personnel assessment and faculty development to program and institutional assessment.

Text Mining, Principle Component Analysis, and Sentiment Analysis

Text mining and statistics are similar in their interest in using mathematical models to define relationships and predict outcomes. Text mining uses the language of statistics to define data and there is no real difference between the two approaches in the common elements of how data is collected and processed. However, there are some fundamental differences. The authors identify major areas of difference as 1) the role of theory, 2) generalizability, 3) hypothesis testing, and 4) level of significance (Zhao & Luan, 2006).

PCA analysis is a particular type of text mining that finds significant groups of words associated with certain characteristics. The goals of PCA analysis is to 1) identify a set of related words in a document, 2) identify clusters of similar groups of documents, 3) to identify clusters with a set of conditions or factors, 4) exploratory

analysis using either structured or unstructured data to discover hidden patterns that might provide useful insights, and 5) to identify frequent item sets and word clusters relating to different factors previously identified (Nisbet, Elder, & Miner, 2009).

One of the central features of text mining is its ability to conduct multiple analyses using several outcomes from the same dataset for the purposes of obtaining the optimal solutions. This approach lends itself to an algorithmic bias. Angus (2003) pointed out that with text mining, every possible report is evaluated, with the most relevant results delivered.

Sentiment analysis operates on a few basic assumptions; first, that there are two main types of textual information that are being looked for: facts and opinions. Most types of text information processing methods are meant to look for factual information. Most statistical surveys that are done for course evaluations are an attempt to convert opinion into a more factually based language that can be more easily synthesized. Sentiment analysis, or opinion mining, allows for the study of opinion, perceptions, and emotions expressed in text. This is of immense importance to colleges and universities because of the amount of written and unstructured text that is available about the institution, programs, departments, instructors, and administrators. One of the chief limitations of the current form of survey is that it will only reveal the answers to questions that the designers think or have the ability to ask about and lacks in the ability to adapt or respond to issues or threats that are not specifically mentioned (Lui, 2010).

In a basic open-structured request for information, the type of questions seen at the bottom of a course survey that might say something like, "Please give any additional comments about the course." The type of information presented will be facts, opinions, targets of opinions, and opinion holders. The subject of the comment is known as the "target object" which can be a product, person, event, organization, department, or topic (Figure 2). It is often represented as a hierarchy of components and subcomponents, with each often associated with a set of attributes.

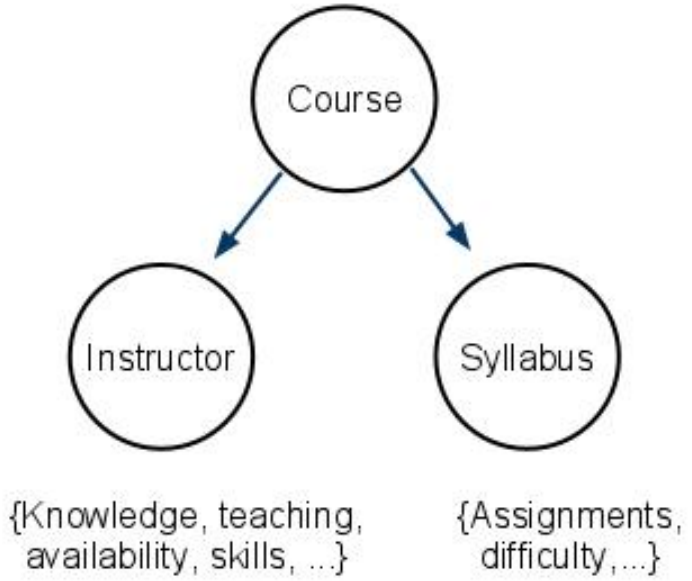

Figure 2. Diagram of a target object (Lui, 2010).

An opinion, in sentiment analysis, is described as O_j, F_{jk}, S_{ijkl}, H, T_l where O_j is a target object, F_{jk} is a feature (a set of attributes) of the O_j, S_{ijkl} is the sentiment value of the opinion holder H_i on a particular feature, F_{jk} of the object O_j at a

particular time t_l. S_{ijkl} is also either positive +ve, negative -ve, or neutral neu; H_i is the opinion holder, T_l is the time when the opinion was expressed (Lui, 2010). The primary objective to sentiment analysis is to apply a structure to the unstructured text. This can be done through discovering and mapping all quintuples, that is, mining the five corresponding pieces of information in each quintuple. Traditional data and visualization tools can then be used to analyze and visualize the results in multiple ways, enabling the use of quantitative and qualitative analysis tools.

There are several levels that sentiment analysis concerns itself with: document level, sentence (or clause) level, and the quintuple level. With document level analysis (in the case of a survey, each open-ended question would be considered a document) each document would be classified based on an overall sentiment expressed by the opinion holder, either positive or negative. There is an important assumption here, however, which is that each document focuses on a single object and contains opinions from a single opinion holder. This could prove problematic with more complex sentiments where the primary opinion holder is expressing the thoughts of another opinion holder (Example: "I told my friend about the assignments that we were doing in class and she, having taken it from another instructor, thought it was ridiculous").

Sentence-level sentiment analysis has one basic task: classify the sentence as either objective or subjective. If the sentence or clause is subjective, it is then further classified as positive or negative. There are a few caveats, however. First, subjective sentences do not necessarily mean that there was a positive +ve or negative -ve

opinion (for example, "I think that I needed this class for my degree"). Second, an objective sentence does not necessarily mean that there is no opinion (for example, "The instructor let the class out early" which may imply a negative opinion).

There are limitations to both the document and sentence level classification because they do not address the complexity of the issues presented. They deal in absolutes and may miss what students truly like or dislike because a positive opinion on an object does not mean that the opinion holder likes everything, just as a negative opinion does not mean that the opinion holder does not like anything. Therefore, it is essential that the text is broken down into quintuples and that all quintuples are found.

This is a difficult problem and involves a variety of steps: 1) named entity extraction (the target object O_j), 2) information extraction (a feature of the object), 3) sentiment determination, 4) information and data extraction (contextual and time stamp), 5) coreference resolution (website = blackboard = LMS), relation extraction (objects are nouns), synonym matching (like, good, great). In commercial settings, it is noted that sentiment analysis is easier to perform because the objects and entities are given and, because of the focus and intent of the review, there is very little noise. Forums and blogs, however are harder because often the objects are not specifically referred to or given and there may be a fair amount of noise. Overall, determining sentiments are easier to decipher while determining the objects and their corresponding attributes tends to be more difficult.

In the context of using sentiment analysis on student comments, the following is instructive in that it shows that, within the context of a course survey, the hardest part of sentiment analysis is already known—that of the object.

Summary

Despite many of the myths, half-truths, and fears surrounding student evaluations, more than 90 years of validity and reliability research that supports the value of student course surveys as a method of data collection on student satisfaction and student perception of teaching effectiveness. Results of the surveys, however, may be influenced by the purpose of the survey. Survey data collected for the purpose of evaluation is perceived differently by both faculty and students than is the data collected primarily for instructional improvement. Timing also has been shown to have a demonstrable effect on the results as surveys conducted in the middle of the semester tend to demonstrate greater student participation and satisfaction that the same survey conducted at the end of the semester as students may feel that they may have an effect on their own course's quality toward the middle of the semester whereas a survey conducted at the end will have little perceived impact on their own experience.

Despite any shortcomings, text mining provides a highly structured way of adding to the data gathered from the Likert scale section of the course surveys, allowing for greater context to be considered in conjunction with the statistical data. Student written comments may provide additional useful and actionable information

beyond what can be learned through the standard Likert scale questions. Without tools such as text mining, the unstructured data presented in the open-ended questions of the course survey instrument is problematic and difficult to access or integrate into program and institutional level statistical inquiry.

In the next chapter, the research design and methodology will be outlined and discussed in relationship to the study's research questions.

CHAPTER III

METHODOLOGY

Introduction

This chapter provides an overview of the research design for this study including the embedded correlational model used to integrate the text mining workflow with the results of a qualitative analysis. The tools used for collection and analysis will be discussed in detail along with descriptions of the population selection and sampling strategies, data collection, and analysis methodology for the research questions. The chapter will conclude with a discussion of methodological limitations and ethical considerations for the study.

Research Design

This study employs the use of an embedded correlational model design (Figure 2). The mixed methods design consists of two distinct lines of inquiry: the quantitative study informed by the qualitative analysis (Creswell, 2003). In one line of inquiry, the data were collected from student survey data, first analyzing the correlation between student responses on the qualitative section of the SACCP surveys and the quantitative section using PCA analysis (Braskamp et al., 1981; Alhija & Fresko, 2009; Luan, Zhao, & Hayek, 2009) and then using factor analysis, and K-Means and Two Step clustering techniques based on proximity or distance techniques to develop a typology to provide an institutional level analysis of the unstructured data (figure 3).

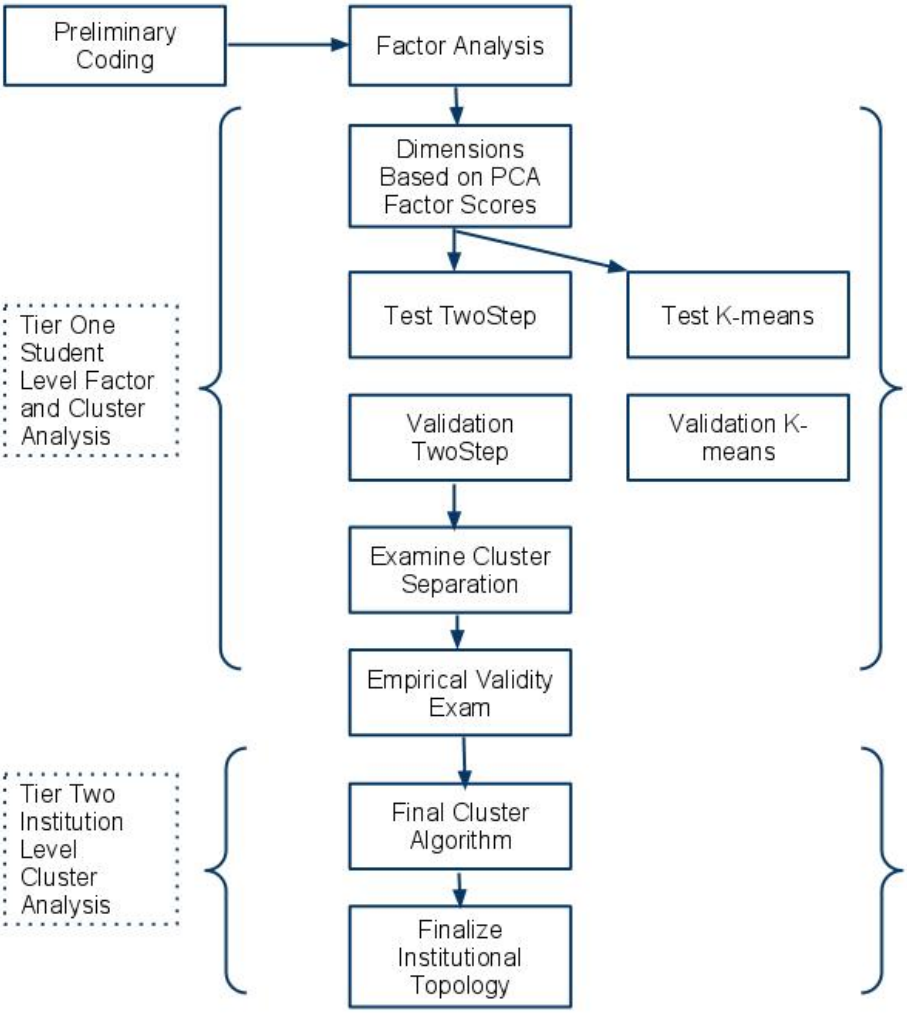

Figure 3. Text mining workflow (Luan, Zhao, & Hayek, 2009).

This process was then revised to address specific concerns regarding the structure and nature of the initial unstructured data and the research questions of this study. Figure 4 details the process that is outlined later in this chapter.

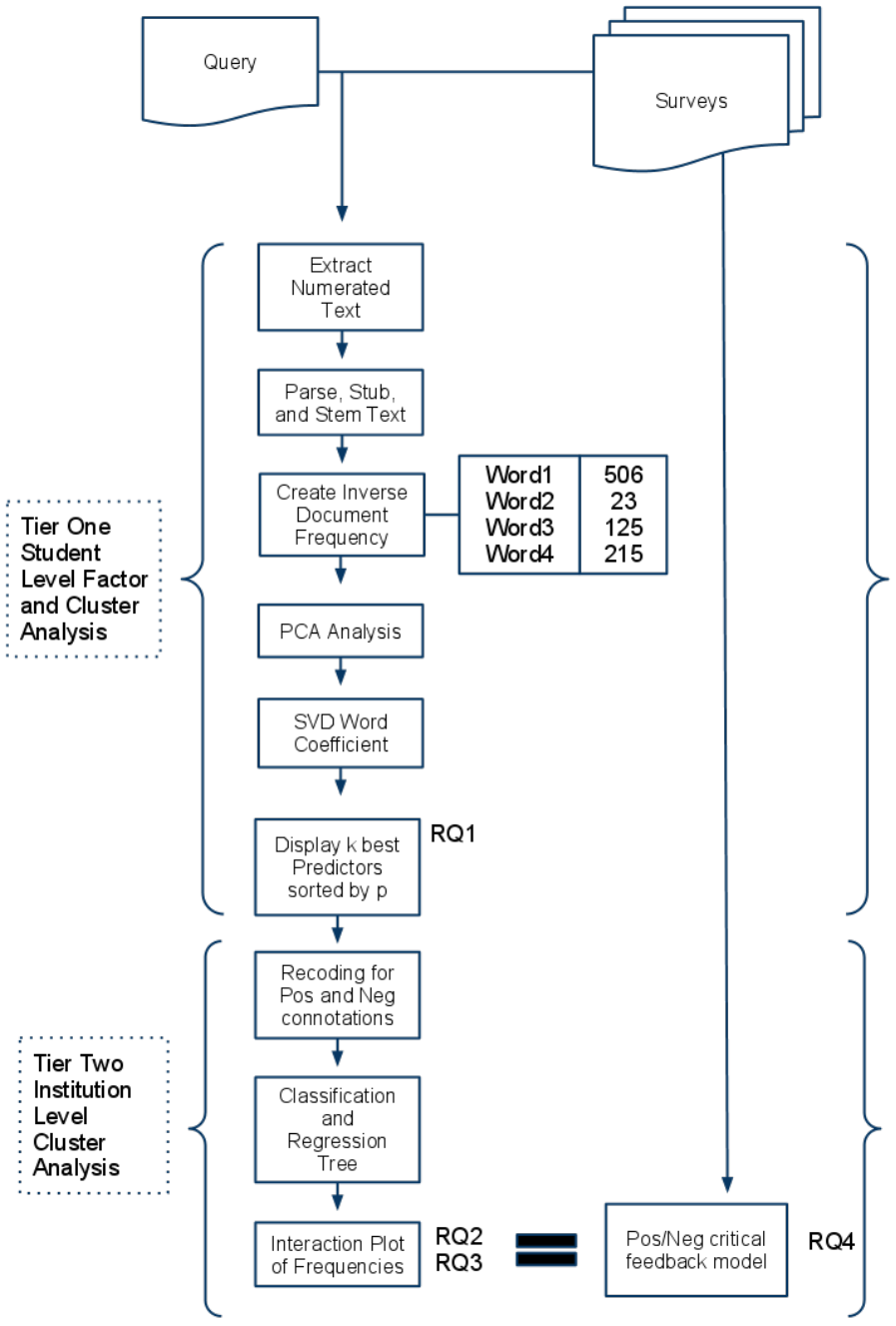

Figure 4. Revised text mining workflow based on Luan, Zhao, and Hayek (2009) and Nisbet, Elder, and Miner (2009).

The second line of reasoning was a qualitative analysis of student comments provided by the SACCP survey which is conducted at the end of every class in the program. The rational for this approach is that the quantitative analysis provides both an overview and a baseline understanding of the relationship between student preparedness and subsequent performance. The qualitative data provides a broader understanding of the challenges and reasoning students give for their performance. The qualitative analysis in the embedded correlational model (Figure 4) is meant to provide a broader grounding and expansion of the discussion provided by the quantitative data (Glaser & Strauss, 1967; Flick, 2006).

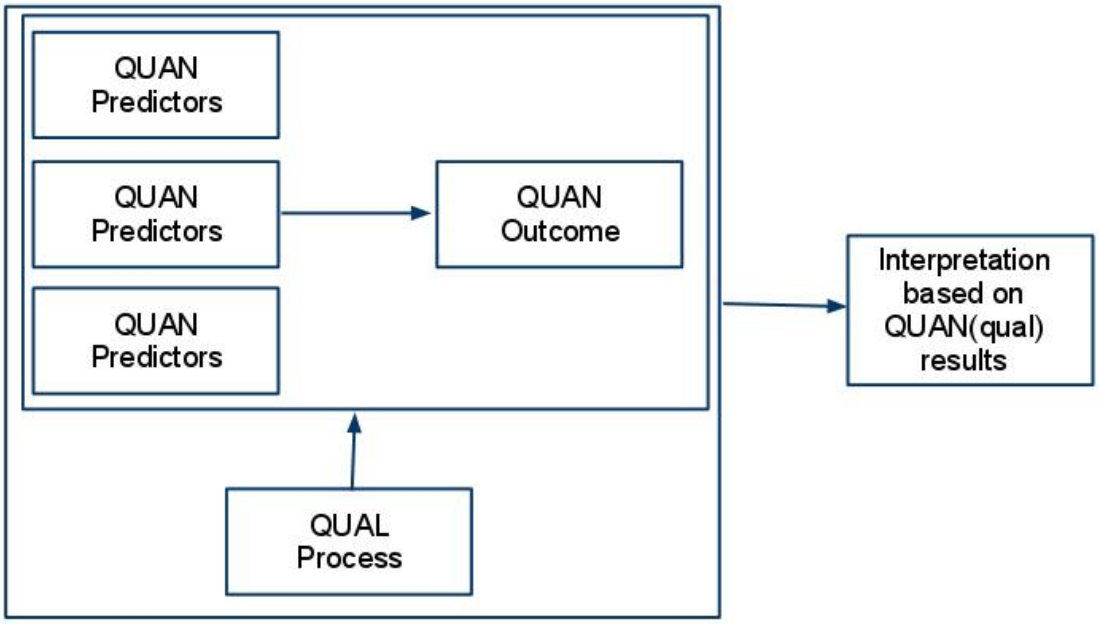

Figure 5. Embedded Correlational Model. (Flick, 2006).

Tools

The software used in this study consists of StatSoft Statistica with TextMiner for the text mining operation, SPSS Version 19 for both description and statistical analysis, Microsoft Excel and Google Docs for input and data organization, and a custom coding database created in Microsoft Access for coding (see Appendix B).

Populations Selection and Sampling Strategies

The data for this study were taken from the end-of-class student survey from the Substance Abuse Counselor Certificate Program (SACCP) from the Center for Professional and Continuing Education (CPCE) at the University of the Pacific in Stockton, California. The CPCE offers a variety of professional and continuing educational programs and classes. The SACC course meets the educational and training requirements from the State of California Department of Alcohol and Drug Programs and was developed by the CPCE in partnership with professionals and experts in the field. The data includes students who participated in the program and were matriculated between 2005 and 2009. This program is offered once a year and operates under a cohort model. Each student admitted to the program takes the same courses in the same order over a six-month period. By the time they complete, they will have filled out up to 14 course surveys regarding their experience with the program.

The data were collected from a four year period from 2005 to 2009, and was taken from paper records on file at the CPCE. The survey is conducted anonymously

at the end of the quarter, so no personally identifiable student information is present on the written records. The sample consists of *N*=835 anonymous student surveys. No gender or other demographic information is available in the sample. The written records were transcribed using a form in Google Docs that closely matched the original form. Accuracy of the transcription was established by comparing the new transcriptions from the student's handwriting with the original summary transcriptions provided by the CPCE (which contain only student comments independent of the quantitative information). The instructor information was excluded, replaced by a random ID number so that the data could remain associated with a particular instructor and course, but was separated from any personally identifiable information such as instructor names or other demographic information. Once transcribed, the data was exported to a csv file and imported for text mining into Statistica, a custom Access database for the coding of the qualitative statements, and then into SPSS for factor and frequency analysis.

Instrumentation

The course evaluations are administered in class by a proctor toward the end of the course. Every course in this program is evaluated every year. The course evaluation is required for completion of the course, so participation is generally 100% of completing students. While there are students who drop out of the program, there are no transfers into the program once a cohort has begun. The instructor is not present during the administering of the survey, and the process is overseen by a

proctor. Students are instructed to rank their assessments and evaluations of a particular course with 16 questions using a Likert scale ordered from 1-Strongly Disagree to 5-Strongly Agree.

The survey is divided into five sections:

- Course organization with four Likert scale questions and space for an open-ended comment
- Instruction with five Likert scale questions with one comment
- Registration procedures with four Likert scale questions with one comment
- Physical Arrangements with three Likert scale questions with one comment
- A demographic section asking how the student heard about the course and the reason for taking the course
- Overall comment section consisting of five open-ended questions.

Unlike many course evaluations which only contain one to three open-ended questions, the SACCP evaluation form contains one open-ended question for each of its three sections, in addition to five open-ended questions at the end of the survey.

Data Collection and Analysis

Based on an analysis of similar studies, this study has adopted a bottom-up and two-fold approach (Figure 3). A factor analysis was conducted and the derived factor scores were later used to conduct a two-tiered cluster analysis. The first tier

clustering is based on student level data where empirical validation of the results was performed and finalized the institutional typology.

Research Question One: Are the student comments of course evaluations aligned to the quantitative portion of the course evaluation instrument?

Hypothesis One: The student comments of course evaluations are aligned to the quantitative portion of the course evaluation instrument.

First, word frequencies were extracted from each of the nine unstructured data variables (Table 1). A Stop-word file was created with common words such as "class," "course," "instructor," and "teacher" in order to exclude common words and terms with low connotative value from being indexed. Additional words were added based on the Inverse Document Frequency (IDF) calculation which provides a dampening of simple word frequencies by using the log function and includes a weighting factor that evaluates a 0 if the word occurs in all documents and assigns a maximum value if the word only occurs in one document. This process results in a transformation of the data with the creation of indices that reflect both relative frequency of word occurrence as well as retaining semantic value within the documents included.

Latent semantic indexing was performed by running a SVD in order to determine underlying dimensions that account for the most common content and meaning from the high valued words that were extracted. A Scree Plot was then run in order to determine the number of singular values that are useful for further analysis.

A SVD Word Coefficient was run to explore the dimensions into which the words were mapped. The top components identified from the Scree Plot were selected and a Scatterplot was generated that created a proximity grouping using Proximity Analysis Components (PAC) that was used to explore the special distribution of values. Labels were applied to the scatterplot to identify the most significant words containing both positive connotations and their relationships to selected variables and then compared to words containing negative connotations as well as their relationships.

Next, each of the significant words identified by the SVD process was added to the original table as variables for additional analysis. This process allows the selection of any words in the scatterplot with either positive or negative connotation. Then, a *k* best predictors sorted by *p* was run in order to create an importance plot to highlight important predictors. This process was repeated multiple times using the structured data variables identified in Table 1. Evidence of *k* best predictors where *p* is the structured data determined whether or not the hypothesis for Research Question 1 could be rejected.

Research Question Two: Are there words and patterns prevalent in the student comments of course evaluations that can classify individual courses on the basis of negative connotations?

Table 1
Variables from SACCP Evaluation Survey

Unstructured Data	Structured Data
Comments on Course	Course Title*
Organization	Course Instructor*
Comments on Instruction	Course goals and objectives were well stated.
Comments on Registration	Course goals and objectives were well met.
Procedures	Course Matched the description published.
Comments on Physical	Resources used in this class were helpful.
Arrangements	The instructor was well prepared.
Needs and Expectations	The instructor was very knowledgeable about the
Meet	subject.
Strongest Feature	Materials were presented logically and sequentially.
Weakest Feature	The instructor expressed his/her ideas clearly.
Additional Courses	The instructor was responsive to questions,
Additional Comments	comments, and needs.
	Registration procedures were convenient and efficient.
	CPCE staff was helpful and courteous during the registration process.
	Registration instructions were clear and easy to understand
	Cancellation policy was explained either verbally or in writing at the time of registration.
	The classroom size was adequate for the number of people enrolled.
	The room's ventilation/heating was adequate.
	Directions to the campus and classroom were clear.

* Excluded from *evidence of k best predictors*

Hypothesis Two: There are words and patterns prevalent in the student comments of course evaluations that can classify individual courses on the basis of negative connotations.

Research Question Three: Are there words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can provide additional information at a program or institutional level?

Hypothesis Three: There are words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can provide additional information at a program or institutional level.

Due to the nature of Research Question 2 and 3, both will be examined together in this section, as the method follows a similar pattern. The predictor list was modified to include all class names. Because the research question is looking at classifying individual courses, the names were excluded from the text corpus; otherwise, they would skew the results by being the best predictors of themselves.

Course Name was selected as the dependent variable and all extracted words identified in the previous analysis were selected as continuous predictor variables, excluding the course names as they may appear in the extracted text.

To corroborate and specifically address both Research Questions 2 and 3, the data matrixes were recoded to add two new variables: NegCon and PosCon. This recoding allows comparative study between courses and collections of courses in

addition to further validation of Research Question 1 as it relates to positive and negative connotations.

From the text corpus, words were identified as being positive, negative, or neutral in connotation. If the inverse document frequency of variables holding negative connotations of words is > 0, the NegCon value for that word was recoded to a value of '1' while all others were given a value of '0'. If the inverse document frequency of variables holding positive connotations of words is > 0, the PosCon value for that word was re-coded as a value of '1' while all others were given a value of '0'.

Table 2

Positive/Negative Critical Feedback Model

	Description
Positive	P: Feedback showing observation and awareness with positive connotations.
Neutral	N_u: Observation with neither positive nor negative connotation.
Negative	N: Feedback showing disapproval or negative connotations.

To conduct the comparative study, an equal number of cases were selected by first running a frequency for all Class Names. The number of random selections for each Class Name was equal to the Class Name with the lowest frequency. A random

sample was then be extracted for each of the Class Names. An Interaction Plot of Frequencies was generated first using Class Names and NegCon as the grouping variables. Next, for Research Question 3, a second Interaction Plot of Frequencies was generated using Class Names and PosCon as the grouping variables. **Research Question Four:** Is there an association between the results of the text mining analysis of the unstructured data and a qualitative analysis of the unstructured data?

Hypothesis Four: There is an association between the results of the text mining analysis of the unstructured data and a qualitative analysis of the unstructured data.

Qualitative data was coded and analyzed to determine themes and trends of student feedback. A codebook was developed that organized the students' written text into the following categories: instruction, course design, and context (procedure, management, and facilities). Additionally, the positive/negative critical feedback model (Table 2) was applied. Looking at both the qualitative and quantitative data in this light, this analysis ultimately provides a context for developing a model to frame questions of student satisfaction in terms of instructional, program, and institutional improvement and review.

Methodological Limitations

The main limitation of this particular approach is that because historical data is the primary source of information in this study, a central voice regarding the student experience is missing: the voice of the teachers. Though it is beyond the

scope of this study, a follow-up study might be warranted where the model is applied to select groups of students with observations and teacher interviews from the classes being analyzed. Additionally, the method by which universities and colleges collect course surveys may need to be expanded to include instructor observations and assessments of the courses and students that they teach. A holistic and complete approach to this problem might ideally include an analysis of student responses, instructor observations, analysis of semester grades, as well as any additional outside observations by administrators or faculty development coaches. Such a model may be logistically or politically prohibitive, but could yield invaluable information over time for both the instructor and the institution.

The researcher makes an assumption that instructor input is significant (which is expected to be illustrated to some degree in both the qualitative and quantitative data) and the current study is looking at specifically to what extent other factors impact the types of comments students make on their evaluations.

An additional limitation to this study is the inclusion of classes from one program at one university. The makeup and nature of the course evaluation instrument, however, is within the parameters of most college and university course evaluation interments so that the principles and methods outlined here should be easily replicated and expanded on in future research at other institutions.

Ethical considerations

The primary ethical considerations taken into account in this study were the exclusion of student and instructor information in the student evaluations. For analytical purposes, instructor names have been replaced by a generic marker such as Instructor A. Fidelity has been maintained so that only one given instructor identified as Instructor A was always referred to as Instructor A, only one instructor was referred to as Instructor B, and so on. Additionally, these substitutions were made within the text corpus so that a student's reference to an instructor was referred to by the label and not the instructor's name.

Summary

This chapter outlined the basic methods of this research and provide an outline for the results in the next section. The data for this research was taken from a single program over a five year period and consists of both Likert scale questions and open-ended questions that cover instructional, course design, and contextual areas.

Using an embedded correlational model design, this study seeks to examine student written comments as unstructured data by using both quantitative methods through the use of text mining, primarily through the use of PCA analysis, K-means, and two step clustering techniques. In a separate line of inquiry, in an effort to both provide a baseline approach as well as to inform the text mining results, the written comments were also coded by hand using more traditional qualitative tools. This

chapter provided a high level overview of the process and methods employed for this research. The results will be explored in greater depth in the following chapter.

CHAPTER IV

RESULTS

Introduction

This chapter explains the organization of the analysis. The first section covers the purpose and design of the research. Next, there is an explanation of the data collection process and an overview of the variables of the instrument along with the encoding of the data into an Access database for coding and interrater reliability. The text mining setup in StatSoft Statistica version 8.0 is then discussed followed by a detailed analysis of findings organized by each of the four guiding research questions.

Overview

The purpose of this research was to determine whether the written, open-ended comments in the course evaluation surveys provided any useful information at the program or institutional levels, to what degree the open-ended responses correlate with the Likert scale questions, and what potential information could be elicited using text mining. Because written comments are unstructured and difficult to quantify, they are often not included in any significant way in course evaluation analysis and are more often used, if at all, as anecdotal evidence that is rarely seen or analyzed beyond the instructor or the instructor's immediate supervisor or the tenure and promotion committee. Text mining, however, provides a way to not only quantify

and analyze unstructured data, but it can also draw out important program and institutional information that may not have been specifically asked for by the closed-end Likert scale questions.

The design for this research included a mixed methods approach in order to first determine if the unstructured data could be restructured into useable organized data that could reveal additional relationships for analysis using text mining techniques. PCA analysis was used to reduce the number of qualitative variables and to detect structure in the relationships between the variables.

The following research questions guided the direction of this study and informed data collection and analysis:

1. Are the student comments of course evaluations aligned to the quantitative portion of the course evaluation instrument?
2. Are there words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can algorithmically classify individual courses on the basis of negative connotations?
3. Are there words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can provide additional information at a program or institutional level?
4. Is there an association between the results of the text mining analysis of the unstructured data and a qualitative analysis of the unstructured data?

Data Collection

A 25 item course evaluation survey was administered at the end of each class in a six-month substance abuse counseling certification program over a five year period from 2005 to 2009. The survey was conducted in a pencil-and-paper format. Eight hundred and thirty-five surveys (*N*=835) were manually entered into a Google Docs spreadsheet. The document was set up with 29 variables (see Table 1), including the eight variables containing the unstructured data. During the transcription process, all references to specific instructors were redacted and replaced with a more generic "the instructor," as per the agreement with the department providing the course evaluations. Once the transcripts were completed, a random sample of 160 records was selected and compared against the original pencil-and-paper evaluations for accuracy. The Likert scale entries displayed an accuracy of 100% while the unstructured text (*N*=2983 words) displayed an accuracy of $> 99\%$ with all of the errors (*N*=23) accounted for by the inadvertent correction of typographical errors in the transcription.

The unstructured data transcribed consisted of *N*=14,242 words with only 64 of the 835 records containing no responses to any of the open-ended questions, giving a 92% response rate to the open-ended questions. This is a relatively high response rate for written comments, which are generally closer to 10% to 50% (Theall & Franklin, 1991; Hardy, 2003; Zimmaro, Gaede, Heikes, Shim, & Lewis, 2006; Alhija & Fresko, 2009). Of the students who wrote comments, the average words per person across all questions was 17 with a *SD* of 16.8 indicating a wide range of response

Table 3

Total Mean and Standard Deviation of Words by Question

Question	Responses	% of Total	Total Words*	Mean**	SD
Comments on Course Organization	99	11.9%	855	8.64	6.80
Comments on Instruction	129	15.4%	1124	8.71	7.12
Comments on Registration Procedures	107	12.8%	935	8.74	7.58
Comments on Physical Arrangements	161	19.3%	1398	8.68	6.37
Needs and Expectations Meet	703	84.2%	2167	3.08	4.16
Strongest Feature	688	82.4%	3328	4.84	4.28
Weakest Feature	379	45.4%	1586	4.18	4.67
Additional Comments	359	43.0%	2849	7.94	6.69

* Calculated in Excel using the formula =LEN(TRIM(A1))-LEN(SUBSTITUTE(TRIM(A1)," ",""))+(LEN(A1)>1)
**Words per entry

lengths from 1 to 50 words. The responses ranged from an average mean of 8.7 words to 3.1 (see Table 3) words per question, with the question, "Does the course meet your needs and expectations?" eliciting the strongest response rate of 84.2% (*N*=703). However, 465 of those responses were single word answers.

In the preliminary review of the students' written comments in each of the eight open-ended questions, an Access database was used to organize the dataset, and each record had between zero and eight written responses. Each response was broken down into meaning phrases with each phrase being assigned one or more content codes as well as a connotation code if the meaning phrase was identified as having either a positive or negative connotation. Initially, 47 codes were created and a sample of 20% (N=170) records were precoded. During the precoding process, the codes were defined to a greater detail and combined to reduce redundancy. The final codebook (see Appendix A) identified three focus areas (instruction, course, and context) and was reduced to 27 content codes. Initially, five connotation codes were identified and coded (positive constructive, positive, negative constructive, negative, and natural). Interrater reliability was conducted with the help of two graduate students in the Educational Leadership (Ed.D.) program. Each was given the codebook along with definitions for each code (see Appendix A). They were then given 100 random statements to code which was then compared to the researcher's original coding. However, the initial comparison failed to rise above a Kappa > .50. The positive constructive and negative constructive code was dropped as it was difficult to define "comment contains constructive criticism" across several readers. While all readers agreed, "This course was excellent" is not a constructive statement. It became more difficult when comparing that to "This class was excellent due to the knowledge and focus of the instructor," which is certainly more constructive, but still arguably lacking in detail. The two codes, positive constructive and negative

constructive, were dropped, along with the "neutral" connotation code. With these changes, the Kappa statistic was performed to determine consistency among coders and the interrater reliability for the coders was found to be Kappa = 0.86 (p <0.001), 95% CI (0.809, 0.907).

For the text mining portion of the study, the dataset was downloaded into an Excel file and imported into Statistica. For the qualitative assessment of the data, the same Excel file was imported into an Access database for coding and an interrater reliability analysis using the Kappa statistic to determine consistency among raters (Landis, 1977).

Text Mining Setup and Analysis

The entire dataset was uploaded into StatSoft Inc.'s Statistica version 8.0 (2008) and was set up using the Text and Data Mining tool. The text mining tool was first applied to all seven open-ended variables. Filters were set to process words in order to include as many cases as possible while excluding likely outliers (see Table 4).

A stop-word file was used in order to exclude words from the analysis that would have little value. The Standard English Stop-Word file included in STATISTICA was applied initially in order to exclude words that are less likely to contain any real value for the analysis. Words in this list include pronouns, verb forms "to be," articles, prepositions, conjunctions, and other common or ambiguous English words. The initial analysis was run using the standard Stop-Word file;

Table 4

Filters Applied to Text Mining Tool in STATISTICA

Filter	Value
Minimum size of word	1
Maximum size of word	25
Minimum size of indexed word	2
Minimum number of vowels	1
Maximum number of consecutive consonants	5
Maximum number of consecutive vowels	4
Maximum number of consecutive same characters	2
Maximum number of consecutive punctuations	1
Minimum percent of files word occurs	1
Maximum percent of files word occurs	100

however, multiple iterations of the analysis were done and the file was modified in order to exclude additional words that contributed little to the analysis (Table 5). A synonym file was added in order to make sure that synonyms (Table 6) were counted together. An initial frequencies list was run multiple times and synonyms were selected as they were found in the results.

Table 5

Stop-Words Added to the Standard File

back, can, could, drug, far, get, given, go, hiv, just, none, nothing, one, put, should, st, treatment, way, will, would, eel, grammar, keep, lot, much, overall, s, t, n, really, take, thing, us

Table 6

Synonyms File Added to Text Mining Analysis

Root	Words
larger:	larger, bigger
deal:	deal, cover, dealt
feel:	feel, felt
info:	info, information
instructor:	instructor, teacher, professor, prof
room:	classroom, room
made:	made, make
people:	people, person
taught:	taught, teach
course:	course, class

The text mining process resulted in the extraction of 111 words from the data set. For the extracted words, the inverse document frequency option was selected in

order to help reflect the specificity of the resulting words through document and word frequencies. This method dampens the simple word frequencies with a log function and includes a weighting factor that evaluates to '0' if a word occurs in all records while evaluating a maximum value if the word occurs only once. This results in an index of frequencies-of-occurrence which includes semantic relationships between records.

Analysis and Evaluation of Findings

Research question one. Are the student comments of course evaluations aligned to the quantitative portion of the course evaluation instrument?

An additive connotation index was created (Table 7) and Pearson correlation coefficients were calculated between the dimensions of student ratings and the dimensions of written comments similar to Braskamp et al, (1981). The findings showed moderately positive correlations between the three dimensions of written comments (instruction, course, and context) and the three categories of student ratings (r =.331 - .459, see table 8). Other factors outside of instruction and course content were unrelated to the instruction ratings; however, the course index was weakly correlated (r =.17) to the context ratings. Written comments on instruction appear to have a moderate correlation to both course and instruction ratings (r =.41 - .46).

The course and instruction index and course and instruction ratings all are weakly to moderately correlated (r =.26 - .46). To determine why instructor and

Table 7
Positive and Negative Connotation by Coded Content Areas of Student Comments

Coded Content Area	No. of Meaning Phrases	Mean % of Students Who Commented by Category	Negative Connotations	Positive Connotations	Connotation Index
Absent	5	0.0060	5		-1.000
Clarity	39	0.0467	14	25	0.282
Competence & Professionalism	38	0.0455	13	24	0.297
Expression of thanks	39	0.0467		39	1.000
General	181	0.2168	5	176	0.945
Helpful, Timely, Feedback & Flexible	61	0.0731	7	54	0.770
Instruction & Presentation	243	0.2910	25	217	0.793
Knowledge & Experience	160	0.1916	2	156	0.975
Multiple Instructors or Guests	19	0.0228	10	9	-0.053
Organization and Focus	70	0.0838	20	50	0.429
Personality	50	0.0599	4	46	0.840
Rigor	12	0.0144	4	8	0.333
Assignments	177	0.2120	29	148	0.672
General	66	0.0790		66	1.000
Materials	136	0.1629	37	99	0.456
Objectives, Organization, and Information	253	0.3030	27	226	0.787
Suggestion	42	0.0503	35	6	-0.707
Time	85	0.1018	62	23	-0.459
Accessibility & Comfort	177	0.2120	153	23	-0.739
Equipment	21	0.0251	20	1	-0.905
Financial Aid	2	0.0024	2		-1.000
General	15	0.0180	5	9	0.286
Registration	26	0.0311	8	18	0.385
Support	17	0.0204	9	8	-0.059
Interaction with other students	126	0.1509	27	98	0.568
Reference to Future	22	0.0263	2	18	0.800
Self Referential	133	0.1593	9	121	0.862

Table 8

Pearson Correlations Between Additive Index Measures of Written Comments and Student Ratings.

Index of Written Connotations	*N*	Student Ratings		
		Course	Instruction	Context
CourseInd	421	.331**	.262**	.171**
InstInd	525	.415**	.459**	-0.01
ContexInd	246	0.10	-0.06	.387**

** Correlation is significant at the 0.01 level (2-tailed).

course ratings and comments are all moderately correlated, a principle component factoring analysis with Varimax rotation and Kaiser normalization was run that resulted in three factors. However, the factors diverged from the student ratings categories on the instrument. While the instrument contained the sections courses, instruction, and context, the factor analysis found little if any difference between the course and instruction sections while separating out registration and physical accommodations in two other factors (see Table 9). While unexpected and inconsistent with the factor analysis established in other studies (Fresko &Nasser, 2001; Alhija & Fresko, 2009) it nevertheless informs the results of the Pearson correlations and allows acceptance of the hypothesis that student comments of course evaluations are moderately aligned to the quantitative portion of the course evaluation instrument.

Table 9

Factor analysis results for student ratings

Item	Course & Instruction	Context	
		Registration	Physical
Course goals and objectives were well stated.	.807	.320	.171
Course goals and objectives were well met.	.838	.276	.124
Course Matched the description published.	.814	.286	.131
Resources used in this class were helpful.	.820	.280	.121
The instructor was well prepared.	.868	.179	.050
The instructor was very knowledgeable about the subject.	.828	.154	.077
Materials were presented logically and sequentially.	.873	.242	.072
The instructor expressed his/her ideas clearly.	.873	.167	.035
The instructor was responsive to questions, comments, and needs.	.806	.167	.067
Registration procedures were convenient and efficient.	.281	.871	.174
CPCE staff was helpful and courteous during the registration process.	.317	.855	.119
Registration instructions were clear and easy to understand	.292	.874	.173
Cancellation policy was explained either verbally or in writing at the time of registration.	.247	.813	.214
The classroom size was adequate for the number of people enrolled.	.065	.109	.890
The room's ventilation/heating was adequate.	.109	.138	.900
Directions to the campus and classroom were clear.	.131	.264	.802
Eigenvalue	8.68	2.45	1.51
% Explained variance	54.26%	15.31%	9.45%
Croncach's α	.96	.93	.86

Research question two. Are there words and patterns prevalent in the unstructured data of student comments of course evaluations that can classify individual courses on the basis of negative connotations?

The NegCon variable was added and recoded to '1' for records where the SVD values of *weak, wasn('t), stairs, small, size, short, screen, room, need, move, little, hot, cold, don('t), didn('t), different,* and *confused* were > 0 or recoding '0' for records where those words contained a value of '0'. Therefore, when the variable NegCon = 0, there is a higher probability of neutral or positive connotations while '1' indicates a probability of negative connotations.

To verify this coding, a qualitative study was conducted where each phrase was coded by the researcher to be either positive (SentNegRev = 0) or negative (SentNegRev =1).

A frequency analysis was conducted between the text mining algorithm variable NegCon and the SentNegRev variable that was coded by hand. The algorithmic process coded *N*=236 records with negative connotations while the hand coded process identified *N*=309 records as having one or more negative statements.

Descriptive statistics were conducted on the NegCon and SentNegRev variables to establish that there was a difference between the quantitative results of those with a negative connotation and those without. In every case, the mean was greater in the group that had no negative connotation statements in the written comments.

Finally, a one-way within-subjects ANOVA was conducted with the factor being the number of methods compared and the dependent variable being the NegCon ($M = .28, SD = .451$) and SentNegRev ($M = .37, SD = .483$) variables. The results for the ANOVA indicated a significant effect, Wilks's $\Lambda = .97$, $F(1, 834) = 29.47$, $p <= .001$, multivariate $\eta^2 = .034$. Follow-up polynomial contrasts indicated a significant linear effect with means increasing $F(1, 3.2) = 29.47$, $p < .001$ partial $\eta^2 = .034$ with the manual coding of SentNegRev over the algorithmic coding of NegCon. The results confirm the observations illustrated in Table 10, of improvement in mean scores by questions when separated into positive and negative categories by manual coding over the algorithmic coding. While separation of records by algorithmic coding had a predictable effect, the manual coding was more accurate. Despite that, the hypothesis that there are words and patterns prevalent in the unstructured data of student comments of course evaluations that can classify individual courses on the basis of negative connotations can be accepted.

Research question three. Are there words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can provide additional information at a program or institutional level?

A SVD was run for the inverse document frequency in order to determine the dimensions of meaning underlying the words that were extracted. The analysis

Table 10

Summary of Mean Scores by Question Separated by SentNegRev and NegCon Variables.

	SentNegRev = 0	NegCon = 0	Total	NegCon = 1	SentNegRev = 1
Q1	4.79	4.75	4.72	4.63	4.60
Q2	4.79	4.73	4.71	4.65	4.56
Q3	4.79	4.74	4.71	4.64	4.57
Q4	4.79	4.74	4.69	4.55	4.52
Q6	4.79	4.73	4.70	4.63	4.56
Q7	4.85	4.80	4.76	4.69	4.62
Q8	4.79	4.71	4.67	4.56	4.46
Q9	4.83	4.76	4.72	4.64	4.54
Q10	4.86	4.81	4.77	4.69	4.63
Q12	4.85	4.81	4.79	4.76	4.70
Q13	4.84	4.80	4.77	4.71	4.67
Q14	4.85	4.82	4.79	4.72	4.69
Q15	4.84	4.80	4.79	4.76	4.70
Q17	4.75	4.73	4.48	3.85	4.02*
Q18	4.68	4.65	4.42	3.84	3.98*
Q19	4.75	4.71	4.61	4.35	4.37*

retuned 29 components for which a Scree Plot was created which determined that components 1 (7.75%), and 2 (4.5%) explain 12.25% of the variance (Figure 6).

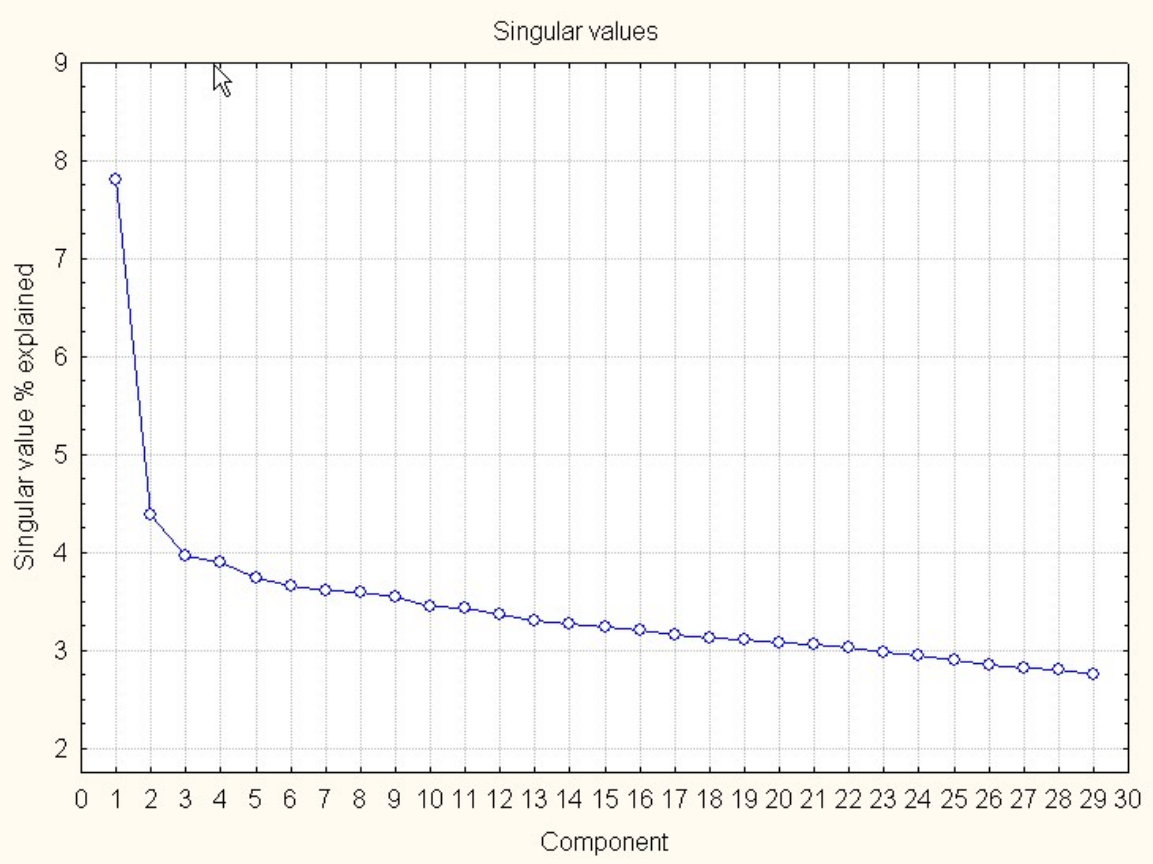

Figure 6. Scree Plot of inverse document frequency singular value decomposition.

A scatterplot was then run on component 1 and component 2 of the SVD Word coefficients in order to visualize the word distribution.

In order to determine the *k* best predictors for the dataset, the SVD Word Coefficients (the 29 SVD components and 111 extracted word frequencies) were re-coded back into the original dataset as 140 new variables. Next, the Feature Section and Variable Screening module was run in order to find the *k* best predictors sorted

by *p*. For the purposes of this analysis, the eight words identified as negative outliers in the SVD Word Coefficients scatterplot were selected (Table 11).

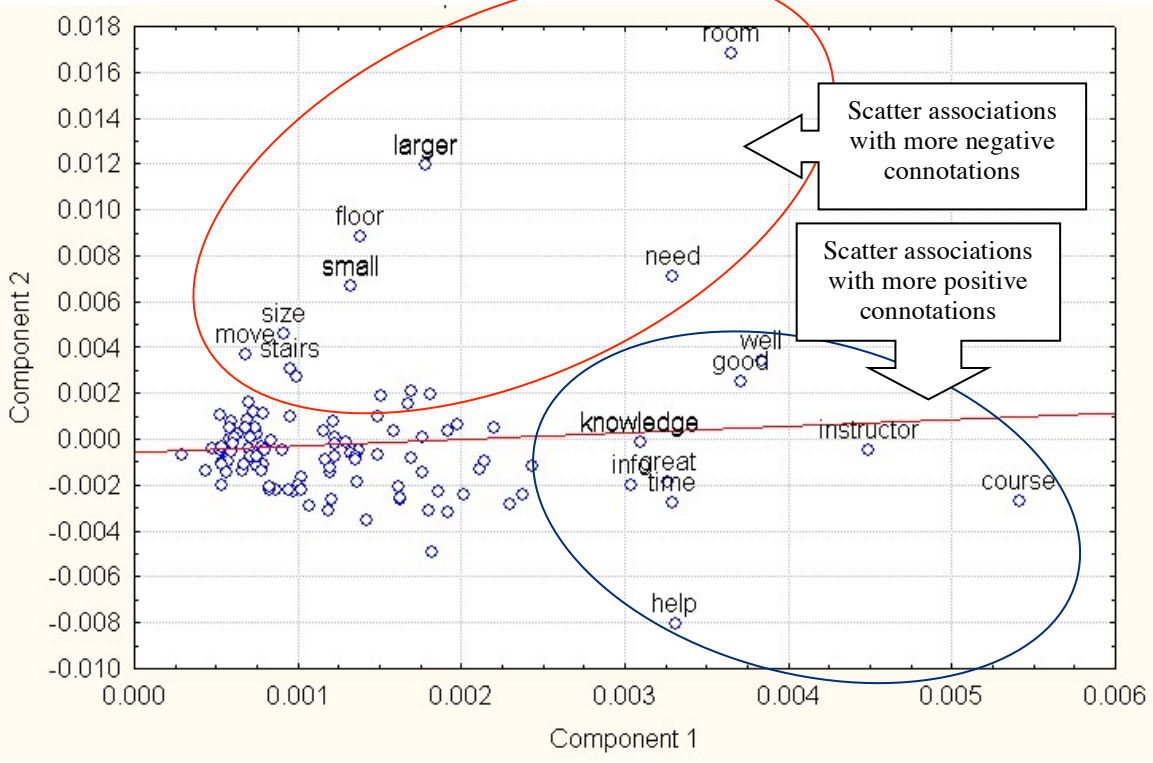

Figure 7. Scatterplot of SVD Word Coefficients with important word associations circled.

The word "room" (.003297, .016833) stands out as a negative outlier so the *k* best predictor was run using "room" as the dependent continuous variable while selecting all the remaining 110 CVD coefficient variables as predictor continuous variables. The results (Table 12) provide a list of *k* best predictors for continuous depended variable of "room" indicating that "larger" ($F = 208.014$, $p < .001$), "floor"

Table 11

SVD Word Coefficients Identified by Scatterplot

Label	Component 1	Component 2
floor	0.001380	0.008831
larger	0.001784	0.011929
move	0.000682	0.003705
need	0.003297	0.007099
room	0.003651	0.016833
size	0.000920	0.004569
small	0.001320	0.006646
stairs	0.000957	0.003076

($F = 145.971, p < .001$) and "small" ($F = 143.677, p < .001$) are primarily predictors of the room being mentioned in the student's written comments. "Size" ($F = 75.266, p < .001$), "move" ($F = 38.607, p < .001$), "cold" ($F = 22.946, p < .001$), and "stairs" ($F = 21.473, p < .001$) are all significant to the p= < .001 level and all appear to indicate possible negative connotations associated to either the size or location of the room, a desire or request to move, or possible difficulties with the stairs.

The word "larger" (0.001380, 0.008831) is interesting as it ranks high in both the SVD Word Coefficients (Table 11), an outlier in the of the Scatterplot SVD Word Coefficients like the word "room" (Figure 7), but is ambiguous in conjunction with

Table 12

Best Predictors for Continuous Dependent Var: Room

Predictor	F-value	p-value
larger	280.0139	0.000000
floor	145.9710	0.000000
small	143.6771	0.000000
need	85.3021	0.000000
size	75.2660	0.000000
move	38.6066	0.000000
cold	22.9456	0.000002
stairs	21.4727	0.000004
absolutely	19.2313	0.000013
first	14.2382	0.000172

the other words identified by the *k* best predictors for "room" as it could be either a request for a larger room, or a reference to a current room as larger after being inconvenienced by a smaller room at an earlier time. Indeed, when the *k* best predictors were again calculated for the word "larger" (Table 13), "need" ($F = 61.155$, $p < .001$), "move" ($F = 36.234$, $p < .001$), and "stairs" ($F = 10.596$, $p < .001$) were significant at the $p = < .001$ level and most likely have negative connotations of needing or desiring to move to a larger room; however, there are the more positive

words "better" ($F = 21.172, p < .001$) which may imply an improvement over an earlier undesired state.

Table 13

Best Predictors for Continuous Dependent Var: Larger

Predictor	F-value	p-value
room	239.6568	0.000000
floor	148.2554	0.000000
need	61.1554	0.000000
move	36.2343	0.000000
better	21.1723	0.000005
due	13.7383	0.000224
organized	12.7190	0.000383
stairs	10.5965	0.001179
well	8.7055	0.003261
job	8.1708	0.004363

The *k* best predictors for "need" also offers an interesting perspective relative to the other words examined (Table 14). Top three words "room" ($F = 84.332, p < .001$), "floor" ($F = 58.556, p < .001$), and "larger" ($F = 53.793, p < .001$), all relate to the classroom environment, presumably the need for more space while "stairs" ($F = 20.981, p < .001$) may indicate an accessibility or mobility issue.

Table 14

Best Predictors for Continuous Dependent Var: Need

Predictor	F-value	p-value
room	84.3318	0.000000
floor	58.5563	0.000000
larger	53.7931	0.000000
time	28.1931	0.000000
well	21.9495	0.000003
stairs	20.9812	0.000005
first	20.7723	0.000006
student	19.7323	0.000010
work	14.7773	0.000130
Don(‘t)	12.7087	0.000385

In order to discover the relationship between rooms and negative connotations, a frequency statistic (see Table 15) was run to determine both the rooms that the course took place in and the number of surveys that were taken in each of the rooms.

Because of the wide disparity of frequencies in room use, a random sample of between 27 and 32 records were selected from each room to adjust for the large variance for the number of surveys completed for each room. A summary frequency was performed on the random selections (Table 16) with cells marked in the

NegCon= 1 column for any instance where the count exceeded the prediction (count $> 10, p = .35$).

Table 15

Frequency Table for Room Number

Room#	Count	Percent
WPC 122	102	12.22
WPC 236	192	22.99
WPC 134	262	31.37
Online	34	4.07
WPC 119	28	3.35
WPC 123	31	3.71
WPC 130	150	17.96
WPC 109	36	4.31

A Pearson Chi-square was conducted to assess whether the room had an effect on the negative connotations of the written comments. The results were significant, $\chi^2(df=7, N=233) = 23.538, p < .001$ with the portion of comments with negative connotations (NegCon = 1, $p = .59$) exceeding the average proportion of the other classrooms of .22 while the proportions of comments without negative connotations (NegCon = 0, $p = .41$) were significantly lower than the mean of the other classrooms ($p = .77$).

Table 16

Frequency of Positive and Negative Connotations by Room Numbers

Room#	NegCon = 0	NegCon =1	Total Records
WPC 122	19	10	29
WPC 236	13	19*	32
WPC 134	20	7	27
OL	24	4	28
WPC 119	21	7	28
WPC 123	24	6	30
WPC 130	20	8	28
WPC 109	27	4	31

*Marked cells have counts > 10

A follow-up test indicated that while the room the course was held in did have a significant relationship on the NegCon = 1 variable, there was not a similar relationship between NegCon =1 and the individual courses χ^2 (df=15, N=233) = 23.538, p = .061 and was not significant at the $p < .05$ level. Another follow-up test looked at the instructor's relationship with the NegCon = 1 variable χ^2 (df=12, N=233) = 25.173, p = .015 and found that it was significant at the $p < .05$ level. The results (Table 13) indicated that while the room did have an significant correlative effect at the $p < .001$ level with the negative connotations in the written comments, and the instructor's effect was significant, only less so at the $p < .05$ level, the courses did not have a significant effect on the negative connotations of the written

comments. In this instance, the analysis suggests that there was a significant issue with Room 236 that affected the overall negative connotations of the comments. Additionally, there is an indication that an improvement in the quality, size, or comfort of the room similarly had a positive effect on the written connotations. As a similar relationship was not found between the NegCon = 1 variable and the courses

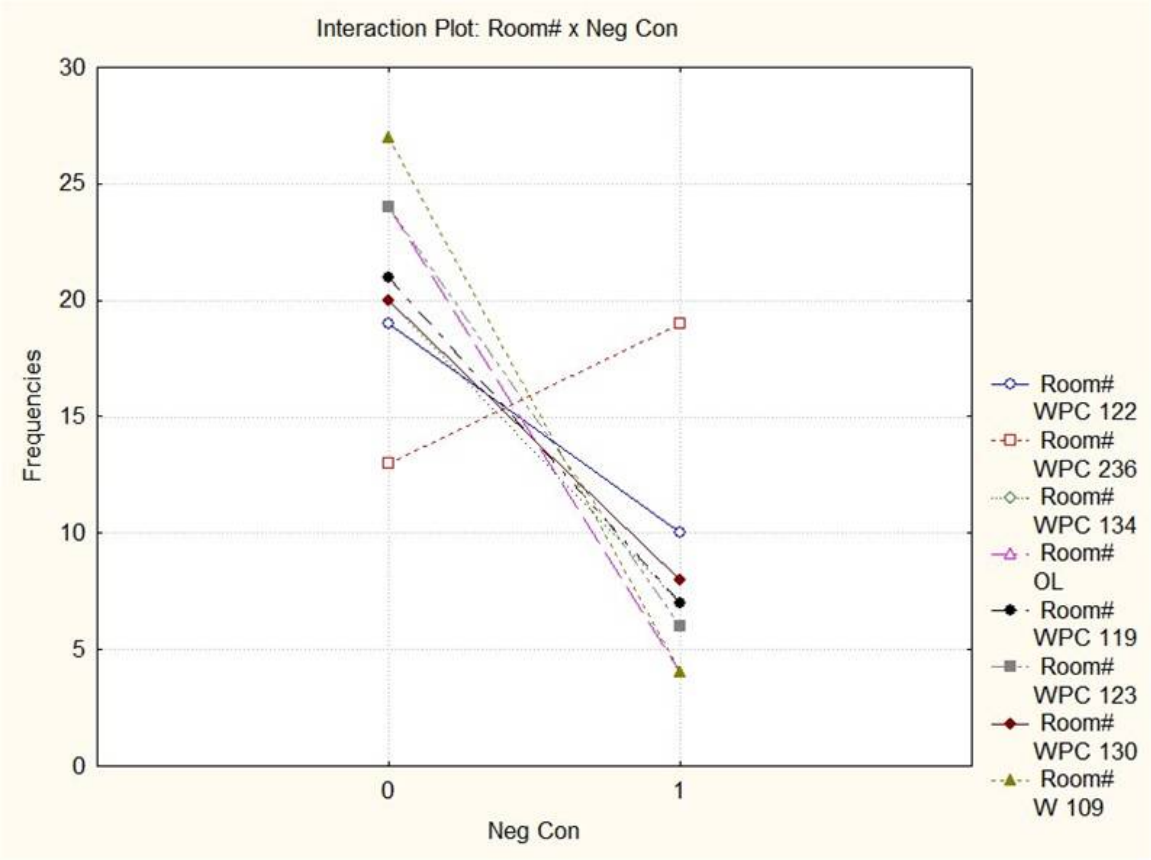

Figure 8. Interaction plot of room number and negative connotations.

in the program, the analysis suggests that there is not a significant enough variance between the courses to create noticeable outliers. A slightly less significance between the instructors and the NegCon = 1 variable, however, indicates that individual instructors have an impact on the overall connotations of the written comments

owing, perhaps, to individual instruction style, methods, personality, or other variables. However, these are differences that the survey instrument is designed to look for, whereas problems with rooms are not specifically built into the instrument. Therefore, the hypothesis that there are words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can provide additional information at a program or institutional level is accepted.

Research question four. Is there an association between the results of the text mining analysis of the unstructured data and a qualitative analysis of the unstructured data?

As part of the data analysis of the unstructured textual responses, the Scatterplot of SVD Word Coefficients (Table 11) showed the word "room" as an outlier and the results discussed as part of Research Question 3 narrowed the focus in this area to specific problems around one room in particular. In order to test the association between the earlier findings based upon a series of algorithmic tests and the findings of a more traditional qualitative approach, the coded secondary content area in the code database indicated that "Accessibility and Comfort" accounted for 153 or 28.7% of total negative fragments ($N = 451$). A simple keyword search of "room" within the "Accessibility and Comfort" code revealed 96 instances ranging from more benign criticisms "The classroom were a little small but manageable" (DocID, 353) to "Needed larger room in a downstairs class for persons with medical or physical limitations" (DocID, 390).

The theme of injuries associated with room selection is prevalent with 13 comments indirectly or vaguely referencing health issues such as "[d]ownstairs would have been better due to most of us having injuries" (DocID, 383) as well as directly referencing vocational rehab and Workman's Compensation with general statements like "[m]ost of the students are on worker's compensation and can not be climbing up and down the stairs" (DocID, 389), "need classroom the first floor. Most of us here have injuries (work related)" (DocID, 393) and "[a]t the beginning of the program the classes were held on the second floor and most of the students were workers comp. and it was difficult, not to mentioned the room size- it was fairly small" (DocID, 2015). A number made implicit references to accessibility with more vague statements such as "too many stairs" (DocID, 345).

Stairs were not the only problem students wrote about regarding their rooms: "[w]e need a larger room" (DocID, 367) and variants were the most common. Primarily, explanations were not given by most, but there were a few that were more descriptive, "The classroom was very big but too many desks were in the class not enough room to move around and too many stairs" (DocID, 345) and "[t]he room is too crowded. We need more space in between desks" (DocID, 350). Other comments indicated frustration of frequent complaints without results: "[w]e need a bigger room it is too crowded in room 236. But do listen to us this time" (DocID, 369).

There were several instances in which the option to move classes was available: once from room 236 to room 224 and another time from room 236 to room

142. In both cases, students referenced either the move, or the contrast between the two rooms, indicating their preference for the latter rooms. Room 224 appears to be better, and indeed several students reference the room in a positive way "[t]he room size was adequate because the class was moved to 224" (DocID, 384) and "[w]e want to stay in Room 224" (DocID, 2341) which was never listed as the room on record for any of the courses in this study, but was a classroom that pleased students when their course was moved there from room 236. Room 142 similarly was not listed as a room in the study, but several classes were moved there resulting in positive comments: "Room was too small. Moved to 142" (DocID, 397) and "142 nice big room" (DocID, 430). In each of these cases, the comments came at the end of a semester where students started out in room 236 and were moved during the course of the semester. These observations are consistent with the conclusion reached above for Research Question 3 where the text mining analysis indicated that some change during the course of a semester was responsible for many of the positive connotation statements by students regarding rooms.

During the text mining, a Scatterplot of SVD Word Coefficients was run (see Table 11) to visualize the word distribution. While some of the words with greater negative connotations have been explored in this study, there was another set of words below the line that was identified as having a greater positive connotation: *info*, *great*, *instructor*, *course*, *time*, and *help*. For the purposes of exploring the association between text mining and a qualitative analysis of the unstructured data,

two words from the scatterplot were chosen as they represent the greatest outliers of the group: “course” and “help.”

The *k* best predictors for continuous dependent variables was run for the word “help” in order to identify both the areas that students where students needed help, and reasons for seeking help in their courses. The word “online” ($F = 47.644, p < .001$) was identified as the best predictor (see table 17) followed by “people” ($F = 37.288, p < .001$), and “question” ($F = 29.293, p < .001$).

Table 17

Best Predictors for Continuous Dependent Var: Help

Predictor	F-value	p-value
online	47.6435	0.000000
people	37.2879	0.000000
question	29.2928	0.000000
understand	22.5236	0.000002
plan	21.9686	0.000003
future	20.8386	0.000006
student	20.7168	0.000006
point	17.3342	0.000035
made	16.8318	0.000045

A keyword search of the coding database shows that for the word "help" there were 64 positive connotations and 10 negative. In relation to the word "online" the following positive statements are recorded, "Helpful to have assignments due on a weekend—good balance between written and online discussion" (DocID, 97) and "I have little experience with online courses, but everyone I have contact with was very helpful. (DocID, 2616). Though both statements use the root word help, they mean it in very different ways: the former expressing a preference for course a design issue, the latter for the relationship between them and others, presumably college staff." The word "people" is similarly positive with four positive connotation statements and no negative statements. In each of these statements, help is used in relationship to the student expanding their perception of other people, "She [the instructor] helped us learn to talk in front of people and about prevention" (DocID, 1362) and "The instructor really helped me see how to look at people from assessment" (DocID, 2294). The word "question" similarly appears with help primarily in positive statements such as "The instructor takes the time to really help answer questions" (DocID, 488) but also contains one negative statement, "I felt the instructor was passive about answering questions and was not prepared to help students who needed more help" (DocID, 162). For the word "understand" ($F = 22.524$, $p < .001$) the statements point to functional areas of improvement, "help to understand ones[sic] self and help others"(DocID, 1679) and "help me get a real understanding about assessment" (DocID, 663).

The *k* best predictors for the dependent variable "course" (see Table 18) shows a simpler, less diverse group of words that are often associated with it in comments. According to the coding database, 30 comments including the word course are negative while 105 are positive. The primary negative statements dealt with time "felt like were rushed though" (DocID, 2660), "Length of the course" (DocID, 2210) and "That is was[sic] a short course" (DocID, 1993), though only one statement specifically used the word time, "Not enough time to enjoy the full scope of the course" (DocID, 1997). "Good" ($F = 55.158$, $p < .001$) and "enjoy" ($F = 50.671$, $p < .001$) both have primarily positive connotations with 18 positive statement and no negative statements for good and 12 positive statements and one negative statement for enjoy. Most statements are general such as "very good course" (DocID, 2576) and "Course goals and objectives were pretty good" (DocID, 119). The word "better" ($F = 27.934$, $p < .001$) seems interesting as it indicates improvement over a period of time or a contrast between two states. When compared to the coded database, we see a contrast between two states, such as "if instructor could be persuaded to teach more courses, then the program would be better" (DocID, 2594) as well as indication of instructor improvement "The instructor was much better prepared for this course" (DocID, 159).

Table 18

Best Predictors for Continuous Dependent Var: Course

Predictor	F-value	p-value
good	55.1579	0.000000
enjoy	50.6707	0.000000
time	31.9808	0.000000
help	31.6069	0.000000
better	27.9339	0.000000
counsel	24.1915	0.000001

Looking at the *k* best predictors of the dependent variable "good" is useful in answering the question of what students find valuable in their experience. From Table 19, we see that the best F-value returned for each of the words is lower than the other words that were examined. This is due partly to the fact that "good" is a word that can be used in a variety of contexts, so while it primarily indicates a positive connotation, it does not necessarily indicate a specific sequence of words that are easily extractable. Good course, good work, good job, and good subject all make sense and are not all that descriptive. In the code database, good accounts for 118 statements with positive connotations while also appearing in five statements with negative connotation. Contrasts between positive words and negative connotations provide interesting insights into the overall responses. In every case, the word good is negative in connection to a suggestion. For example, "I know we had a lot to

cover, but more class discussion/ interaction would have been good" (DocID, 235), "Good grammar review should not be 32 hours" (DocID, 921), and "more class discussion would have been good" (DocID, 1916).

Table 19

Best Predictors for Continuous Dependent Var: Good

Predictor	F-value	p-value
course	28.7064	0.000000
counsel	27.8331	0.000000
work	18.3517	0.000021
know	18.3344	0.000021
job	14.8276	0.000127
explain	10.4616	0.001267
subject	9.9490	0.001667

One weakness of this approach is that the text mining algorithm as set up in this study would not have predicted the word "good" to be used with negative connotations; however, the statements "not" and "would have been," which would have been seen as negative, were not identified by the *k* best predictors test.

Summary

This section followed four major guiding questions in an effort to examine whether text mining may prove to be a useful extension for examining general trends

at a university or in a program in addition to the statistical examination of the quantitative data that comprises the majority of most courses surveys. The process of text mining involves taking unstructured data and preparing it in a way that statistical tests can be performed on the entire dataset so that major trends can be easily seen or identified quickly without having to manually read thousands of words.

The first question, asking if the student comments of course evaluations aligned to the quantitative portion of the course evaluation instrument, found that there were moderately positive correlations between the three categories of written comments (context, course, and instruction) and the corresponding student ratings in each section. Factor analysis, however, showed little distinction between two of the categories, "course" and "instruction" while showing two distinct factors within the "context" sections which is a weakness of the instrument, not necessarily the method.

The second guiding question that looked at words and patterns in the unstructured data that could algorithmically classify individual courses on the basis of negative connotations was supported. While the particular method employed in this study was not as successful in identifying positive and negative connotations as the hand coded method, the results showed a predictable effect of the algorithmic approach. Manual coding was more accurate, but the algorithmic approach was close.

The third guiding question looked at the words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can provide additional information at a program or institutional level.

Using a SVD test, a scatterplot identified a variety of associated words that illustrated an issue that occurred regarding rooms and accessibility. While the Likert portion of the surveys showed a drop in satisfaction, the nature and extent of the issue (as well as possible ADA issues) were clearly illustrated by the text mining procedures. In this area, there is clearly an ability to find a facilities issue that would not have necessarily been evident otherwise.

Finally, the question of an association between the results of the text mining analysis of the unstructured data and a qualitative analysis of the unstructured data show that there is a clear association, but using the methods outlined in this study could, absent a clear issue like the room accessibility issue, be considered overly general and vague without the addition of an underlying structure derived by manually coding the unstructured text. Despite that, such an underlying structure incorporated into the algorithm should be the beginning point of future research. Text mining is a process that improves over time as it "learns" and this research represents a step toward that end.

Ultimately, text mining represents a new way to approach evaluation and assessment at the university. Text mining makes available large amounts of data that had been previously difficult and time consuming to parse and analyze. The next chapter will look at some of the promises and pitfalls of this approach, from the possibility of institutional evaluations conducted in real time from multiple sources of input to the trickier question of privacy where algorithmic analysis of text and other

data now considered public could potentially reveal far more than the respondent intends.

CHAPTER V

DISCUSSION AND RECOMMENDATIONS

Introduction

Most colleges conduct some form of course evaluation as a way to provide feedback to instructors and departments regarding the perceptions of students taking these courses (Alhija & Fresko, 2009). After almost a century, course evaluation is a pervasive practice within higher education and there has been extensive research and inquiry into the validity and reliability of using such methods for feedback (Renaud & Murray, 2005). This feedback, however, raises concerns on the part of faculty (Hodges & Stanton, 2007). While providing quantifiable data on a variety of questions, often tailored to the specific needs of an institution, Abrami et al. (2007) point out that ultimately, such ratings look at one primary thing: student satisfaction with teaching. Unfortunately, Coleman and McKeachie (1981) note the "pernicious" tendency for such instruments to put pressure on faculty to become popular in addition to the fact that it may implicitly contribute to a simplistic view that higher rated instructors positively affect institutional offerings with the corollary that lower ranked instructors may not have as positive affect. This simplistic and overgeneralized sentiment is problematic because such instruments may be used by administrators to mean more than just student satisfaction in an effort to find some way to simplify assessments of instructional job performance.

The Likert portion of the student course survey is well studied, in part because it is easy to quantify, despite its limitations. In contrast, the written portion of the course evaluation (an element that is common to most course evaluation instruments) is not as easily studied nor is it easily quantified. Written comments are primarily local in the sense that they are meant to provide insight or fill in the gaps for the instructor or other interested party such as the instructor's dean or members of the faculty's tenure and promotion committee.

Sheehan and DuPrey (1999) found a link between student evaluations and the comments that were written in the open-ended portion of the survey, but acknowledged that despite positive comments out-numbering negative comments almost two to one, the written comments were also a source of anxiety for faculty with fears of comments that were unjustified, uncritical, or cruel and that such comments might have undue influence on tenure and promotion.

One of the limitations in collecting the thoughts and impressions of students is the observation that much of what students write is not useful due to the fact that their observations, if they contain any detail at all, are generally self-interested and uncritical. Even when many of them do have important observations, they fall short in giving information or details that would make their comments truly helpful. Despite this, there are issues that surface such as room temperature, accessibility for veterans and vocational rehabilitation students, as well as some other structural issues such as instructors that leave early, no-shows, and problems with team teaching,

which may have importance at the department or institutional level. Despite this, there are a few students who provide critical detail that can be deemed helpful.

Rather than placing the blame on students being uncritical, educators should ask if there are better ways to prepare students to be observers. Students should be taught how to assess their own education and to make valid judgments using a critical language. This might be difficult for some instructors who may fear this oversight; however, the usefulness of student observations will increase only as students are taught how to observe and make valid comments. Despite this, the real importance of using text mining for the analysis of student written comments, ultimately, is not diminished by the uncritical comments. While the statements such as "very organized" (DocID, 10) and "I need more structure to force me to study more" (DocID, 708) may not seem very helpful from the context of a single class, 835 students providing 2,561 comments does create some clear trends regarding what they find as both engaging and frustrating. So while "[s]ome organization problems" (DocID, 2158) may seem frustratingly vague, within the context of *N*=20 (4.4%) negative comments and *N*=50 (2.4%) positive comments directly dealing with focus and organization, we see emerging themes of what students value and notice.

Hodges and Stanton (2007) agree that student comments, when abstracted out over a large population, may actually reveal both issues and preferences that are common to students and could actually provide insight regarding effective teaching practices. While one criticism of student written comments is that students are not trained observers (Alhija & Fresko, 2009), when assessed collectively through

processes like text mining, the open-ended nature of the questions allows for a broader range of feedback despite the limitations of low return rates, short comments, irrelevant statements, and off-topic remarks. What it shows is a broader picture of what students find important. Chen and Chen (2009) further note that such a collection of data is not only useful for reviewing specific courses or formative assessment on the part of instructors, but may contain key institutional and program data that would be useful in the institutional review process.

Unstructured data, however, is unsuited to traditional statistical examination and, until recently, could only be studied using traditional qualitative methods, which severely limits the number of students and classes that can be examined (Lin et al., 1984). Yet since the early 1990s, with the exponential advancement of computer processing power, more sophisticated methods of textual analysis such as text mining have been developed in order to deal with large amounts of unstructured data that would otherwise not be as useful.

This research, as stated in Chapter I, has two basic aims: 1) to determine if there is a quantifiable correlation between the statistical portion of the course evaluation responses and the written responses using Principle Component Analysis (PAC), and 2) determine whether or not the unstructured data reveals actionable information at the program or institutional level that could not be determined strictly from the Likert scale responses.

For both of these broad goals, the data presented in Chapter IV fulfill both of these aims. While more detail will be given during the discussion of each of the four

research questions, it should be noted that text mining is different from statistical analysis in the fact that, unlike traditional statistical methods where a generalization can be made from a smaller statistical sample, text mining results get better the more data that is processed. Ideally, a text mining system learns over time and its strength becomes the fact that data can be both unstructured and unlimited. This research is far from unlimited, however, and represents only a small early step.

This chapter will look at some of the issues surrounding each of the research questions guiding this examination. Next, it will examine some of the recommendations for improving results, both algorithmically as well as for improving student participation and possible redesigning of the interment. Finally, it will discuss suggestions for future research.

Text mining presents an interesting challenge to higher education—not only in the information that can be teased out of a massive amount of unstructured data, and the possible messages it may be telling us, but also in dealing with the nature of privacy and what can be revealed in such text as well as what it means for leaders and how they might effectively use such data and avoid abusing the instruments or misinterpreting what the data has to say.

Findings and Interpretations

The original survey instrument used in the program selected for the study was atypical of the standard course evaluation in that it had eight open-ended questions (see Table 3) rather than the typical two to three (Abrami et al., 2007; Sheehan &

DuPrey, 1999; Abbott et al., 1990). For the most part, the student responses were examined collectively without considering the context of the question except where comments were vague. In most cases, context for a comment was provided by the students, such as "Assignments helped us really get going an[sic] planning internship, for interviewing, etc." (DocID, 1602). However, there were instances where vague comments such as "Group discussion" (DocID, 1539, 1534,1578) and "homework" (DocID, 1660, 1935) provided no context, but were coded as Course Design: Assignments and Positive due to the fact that they were written in response to the question " What did you feel was the strongest feature of this course?"

These vague, uncritical, and short responses appeared, ironically, in the more narrow questions asking for specific responses. The question "What did you feel was the strongest feature" and its companion question, "What did you feel was the weakest feature of this course" had an average of 4.84 and 4.18 words respectively. The most direct question, "Did this course fulfill your needs and meet your expectations?" had an average of just 3.08 words per response. Yet both the direct question about meeting expectations (*N*=703, 84.2%) and the strongest feature (*N*=688, 82.4%) had the most responses. The weakest feature had a much lower response rate (*N*=379, 45.4%) which is in keeping with the overall tendency for written comments to be more positive than negative. Sheehan and DuPrey (1999) found in their study that positive comments outnumbered negative comments by 2 to 1 and Braskamp, Ory, & Pieper (1981) found the ratio to be 3 to 1, whereas the results of this study showed the ration closer to 5 to 1. The other open-ended

questions were more open-ended with far broader directives such as comments on course organization, instruction, registration procedures, physical arrangements, and additional comments and garnered higher word averages per entry (between 7.94 to 8.74 per comment), though the actual number of responses were fewer (comments on course organization, instruction, registration procedures, physical arrangements ranging from 99 (11.9%) to 161 (19.3%) responses while the section "Additional comments" garnered a fair 359 (43%) number of responses.

One general conclusion that can be made is that there is a tradeoff in the design of the type of questions asked to elicit written responses. The more narrow and constrained the question is, such as, "Did this course meet your expectations?" the shorter the response. In the case of meeting expectations, the overwhelming tendency was to produce one or two word answers, but as the answers became shorter, and less detail is required, the response rate rises significantly. In other broader comment sections, where students are not asked specific questions, the number of words in the response increases significantly, presumably because more detail may be required to relate the context of the statement, but the tradeoff is that there are far fewer responses, dropping from 84% of respondents down to 11% to 19%. In this particular data set, the one exception to this trend was the "Additional Comments" section that, while holding true to a larger word count of 7.94 per response like the other open-ended question, had significantly higher response rates (*N*=359, 43%).

This is important data to consider in efforts to restructure the open-ended questions in the course evaluation instrument as it demonstrates a clear need for at least one truly open-ended question, such as the "Additional comments," yet for the other questions, the more context that is placed within the question, the more responses are encouraged, but the less detail will most likely be provided.

Research question one. Are the student comments of course evaluations aligned to the quantitative portion of the course evaluation instrument? The results of the Pearson correlations between additive index measures of written comments and student ratings (see Table 8) show that there is a significant relationship between the Likert scale responses between the three main instrument divisions, course, instruction, and context, and the corresponding coded written connotations. The relationship between the student ratings in the course section and the written comments were moderately correlated (N=421, r = .331, p < .001) while the relationship between student ratings in the instruction section and written comments coded as relating to instruction were moderately correlated (N=525, r = .459, p < .001) and the relationship between the ratings on context (everything else excluding course or instruction) and similarly coded written comments were moderately correlated (N=246, r = .387, p < .001). These relationships are important insofar as they establish a general relationship between the nature and attitude of the written comments and the corresponding attitudes gleaned from the Likert scale questions. The problem, of course, is that these indicate only a very basic relationship. When a student makes positive statements, then he or she may be more likely to rate aspects

of the similar category higher. Likewise, a student who writes comments with negative connotations may be more likely to rate items in that category lower as well. There are a few points to consider as to why these items are only moderately correlated rather than strongly correlated, which would make more sense. Why would a student give poor ratings on one or more items and then write positive comments in the open-ended segments, or conversely give high ratings in a category while writing a more critical response in the open-ended section? There are a number of reasons why this may happen.

First, the written comment may be outside the scope of the Likert scale questions. The Likert scale questions are narrowly defined and ask for specific reaction in a broad overview; they are a request for a perception at that moment in time. Overall, the student may have a particular impression of the item in question, however, may expand a more nuanced response in a written follow-up. Second, there is the possibility that some of the written responses may reflect either positively or negatively on areas or issues that are not directly addressed by the Likert scale questions, but that still fit within the general category. Third, there may be a psychological component at play that may allow a student to temper an overly critical assessment in the Likert portion with statements that are more positive, or conversely mute a string of 5s and 4s in the Likert assessment with more critical assessments in the written portion. A student might, conversely, temper a negative Likert assessment with a more positive written statement. Lastly, there are errors on the part of the student where, for example, a student mistakenly enters 1 or 2 (Strongly disagree and

disagree) across all the Likert scale questions but then writes very positive statements in the open-ended section, thus indicating a high likelihood (but not with certainty) that the student may have misunderstood the scale response. For a leader or administrator, the primary consideration is the understanding that while the sections are related, neither side tells a complete story by itself.

Research question two. Are there words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can algorithmically classify individual courses on the basis of negative connotations? There are two basic approaches to systematic analysis of unstructured written data: algorithmic and qualitative. Algorithmic refers to the variety of approaches from sentiment analysis to data and text mining where the processes are based on instructions and procedures, but are largely independent whereas qualitative depends on a human agent and human judgment for analysis. In order to determine the negative connotations present within the unstructured text, the dataset was hand coded for negative statements ($N = 451$ statements from 309 respondents). From these statements, certain words appeared more frequently than they did in positive or neutral statements such as *weak, need, confused, didn't*. Other words, while not necessarily negative, tended to show up in negative contexts within the confines of student feedback. A word such as *stairs*, "[i]t was difficult to go up the stairs" (DocID, 432) primarily referenced accessibility issues. The word *short*, a reference to time, most often referred to having too much material for time provided, indicating a possible course design issue. Then there are words like *time* which is not

as narrowly defined and had both positive connotations such as, "I feel that the instruction was a very good teacher and took time to answer what ou[sic] wanted to know" (DocID, 2359) or negative connotations like, "There were a few times that I called the instructor to ask questions and I never got a call back" (DocID, 23) and therefore cannot be easily encoded into the algorithm without significant complexity.

Once the negative connotations were recoded for both the algorithm (NegCon) and hand coded (SentNegRev), the results were compared (see Table 10), and are both interesting and encouraging. Because there is a relationship between the written responses and the Likert results (see above) it was determined that if the words could be classified into positive and negative connotations, then there should be a similar reflection in the Likert responses. For this to be true, then whatever the mean Likert scale score is on any given item, the mean of the Likert responses for surveys containing statements with negative connotations (either NegCon or SentNegRev = 1) should be lower than the original mean while the mean of the Likert responses for surveys containing statements with positive or natural connotations (either NegCon or SentNegRev = 0) should be higher. Indeed, this is the case (See Table 10). Question One, which states "course goals and objectives were well stated," for example, has a mean of 4.72. The algorithmic process effectively separated the feedback so that those identified as positive had a mean of 4.75, 0.03 greater than the original mean. The group identified algorithmically as having negative connotations had a mean of 4.63, 0.09 less than the original mean. For those identified through the hand coded process, the results were even greater with a mean of 4.79 for the positive

connotations, 0.04 greater than the algorithmic positive group and 0.7 greater than the overall mean. On the other side, the group hand coded as negative had a mean of 4.60, 0.03 less than the algorithmically separated group and 0.12 less than the original mean. This pattern holds true for every question (see Figure 9) in the survey with the exception of the last three questions where there is a reversal in the negative connotations where the algorithmic assessments appear to be more accurate than the hand coded.

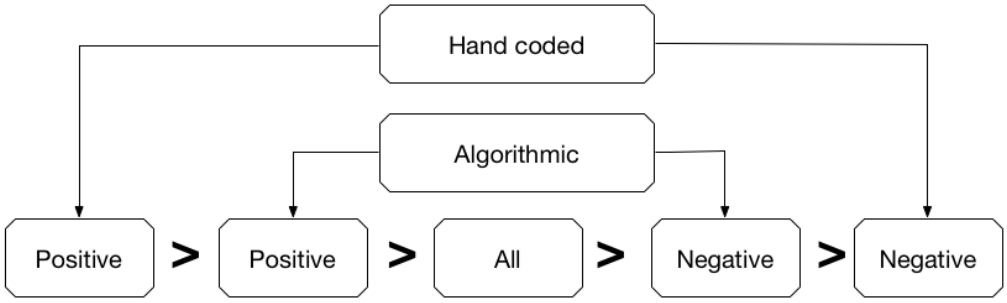

Figure 9. Illustration of results of means from hand coded and algorithmic coding of negative and positive connotations of written statements.

The important thing to note is not necessarily that the hand coded version is more accurate than the algorithmic approach, but that such results were achieved with a relatively minor sorting procedure that could be improved over time with more advanced sentiment analysis algorithms. This analysis was the result of an initial manual process by which the original words and phrases were selected to determine a high likelihood of negative sentiment. For future iterations, these results should be analyzed to a greater degree with the goal of improving the algorithm through a greater understanding of the nuance between word relationships that was not possible

in only one iteration of this process. Next steps for improvement would be an in-depth analysis of false-positives and false-negatives in an effort to refine the instructions set to take into account nuances that cannot be determined by only words with clear negative connotations.

Research question three. Are there words and patterns prevalent in the unstructured data of the student comments section of the entire data set of course evaluations that can provide additional information at a program or institutional level? While the primary areas of inquiry for the course evaluations are generally instruction and course design, many also include specific questions dealing with contextual areas such as registration and facilities. Despite this, such questions are quite specific in order to extract a quantifiable result. For example, a question in the survey examined in this study asks students to rate from excellent to poor, the statement "The classroom size was adequate for the number of people enrolled." The mean response to that statement for the five-year period was 4.48, which tells very little (considering that $N = 835$) other than the fact that it was the second lowest mean on the survey. Even if the data were narrowed to only those courses that showed very low scores in this area, there is little that can be teased from the data other than perhaps the course was overenrolled or the class was uncomfortable. The nature of the question hints at an issue, but lacks the detail to give context or reasons.

Based only on written comments, Figures 7 and 8 show a starker contrast. While an analysis of the Likert scale questions could pinpoint potential problems of room size (because that was the specific question that was asked), text mining, even

without the help of the Likert scale data, is able to pinpoint a room that had consistent problems over a period of time. Best predictor analysis detailed in the previous chapter (Tables 12-14) further suggests that the problem lies not only with classroom size, but with location and accessibility. Using frequency analysis of negative connotations filtered by room number then clearly defines a problem with a particular room, thus allowing for drilling down into the content and context of the negative statements such as "Need class on the first floor and one that's much larger" (DocID, 378) or "From bad to worse - stairs everywhere - handicapped people including I w/o wheelchair and no elevator" (DocID, 344), a potential red flag for facilities management.

Research question four. Is there an association between the results of the text mining analysis of the unstructured data and a qualitative analysis of the unstructured data? As this research is designed, the results of the text mining and the qualitative analysis suggest roughly the same thing in different ways. The primary difference for text mining, as it is conducted here, is that the sorting of data requires analysis, as much depends on the relationship of one word to a group of surrounding words. The qualitative process is easier to approach, as it deals with the interpretation of phrases rather than the more abstract notions of word proximity; however, there are some tradeoffs.

Once all of the data was transcribed into digital form, the qualitative portion of the analysis took several weeks to format and code in order to begin extracting useful patterns. By contrast, the text mining analysis was done in the space of a

single afternoon. The tests were done over a period of a few weeks, in order to hone the order and revise the methodology and update and adjust various aspects of the Stop Word file and the Synonym file as well as the process for determining negative connotations. Once the setup was complete (described in detail at the beginning of Chapter IV) the actual process of running the tests took very little time.

These two methods, while very different in approach, represent both a trade-off as well as a necessary relationship. It is a trade off due to the fact that, as outlined in this research and given that the results are similar in a broad sense, there is a choice between time and detail. Text mining has the virtue of being nearly instantaneous and possibly automated, a core feature if student written comments were to ever be used for data analysis at an institutional level on any consistent basis. What it makes up for in ability to parse large quantities of data, it loses in specificity and detail. Though that detail is still present in the original source data, whether an issue or detail is noticed or not depends upon how well the algorithm is written. The process used in this research was set to ignore relationships that exist <1% of the time. The practical purpose of this is that the process effectively ignores the outliers and puts more weight upon relationships that are repeated multiple times. The focus of the text mining approach is based on extracting relationships out of thousands, if not millions or billions, of individual units of data. This data does not need to be organized in the traditional sense. It does not need to be structured or otherwise processed in order to return meaningful results.

In contrast, a qualitative approach results in a closer level of detail, but represents a significant investment in time and human participation for a relatively small sample. The true cost of this method is the ability to systematize and abstract the process so as to engage in continual collection and assessment. Yet, as the results show, a manual qualitative approach is important for several reasons. First, as is the case in this research, it provides a baseline by which to compare the automated results and then to test both assumptions and interpretations made from those results. Second, to improve the automated results, the qualitative data must be further analyzed in order to improve the results of the text mining (see Figure 10). Additional focus on sentiment analysis and natural language parsing will serve to improve results in future iterations.

Recommendations

Recommendations fall into two basic categories: the collection of the data and how the data is used.

Collection. The primary use of the data captured from most course evaluations is for analysis of specific courses and programs. While some colleges and universities have standardized instruments, some do not, which makes any larger meta-analysis harder to achieve. The real advantage to the institution in using written comments is that the format of the collection no longer limits the ability to conduct analysis on the text.

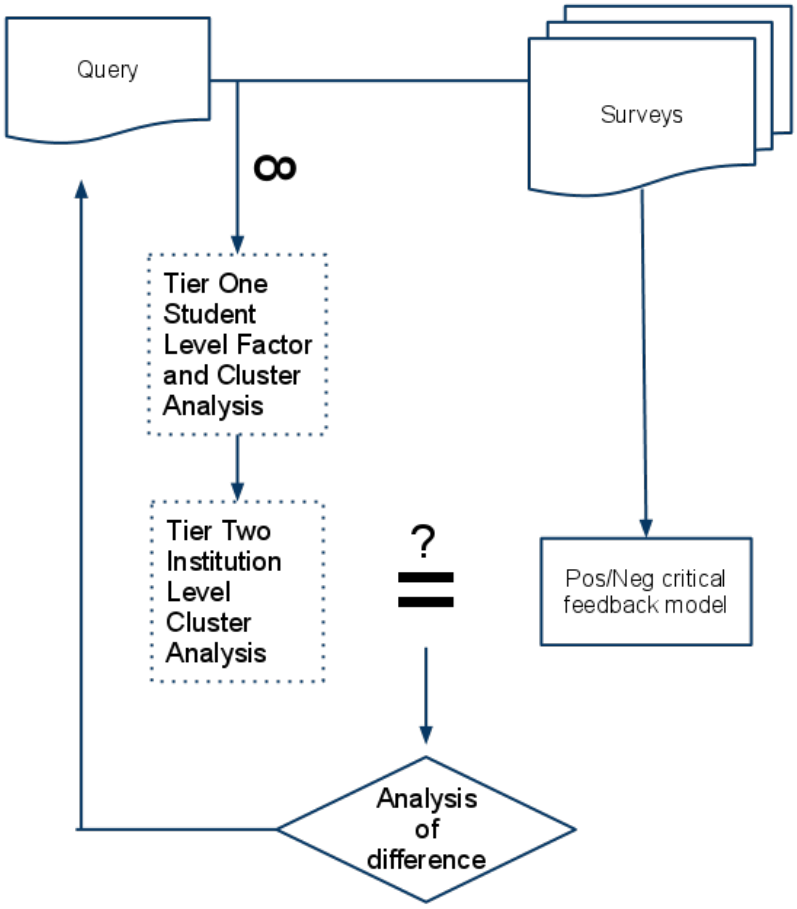

Figure 10. Text mining workflow feedback for improved results.

Collection may be conducted by traditional means with pencil-and-paper, or captured using an online form, or even using a less restrictive format such as the microblogging format best exemplified by Twitter or other electronic means of capturing sentiments and observations. Whether using an algorithmic text mining approach or a more conventional, albeit slower, qualitative analysis, the data that is produced is quite rich, but in a way that is very different from the more statistical analysis of the Likert scale questions. Simply, use of the text mining procedure in

conjunction with the externally derived structure in the code database does indeed reveal a broader and more complete picture of an overall analysis of student written comments.

As Abrami et al. (2007) show, considerations of time, place, and purpose of evaluations are also important. The type of responses and the quality of those responses seem to be determined by a number of factors: the time of the semester or quarter that the surveys are conducted, the stated purpose of the survey, and the evidence that such information leads to either action or acknowledgement.

End of course evaluations, for example, provide little opportunity for concerns to be addressed and even legitimate concerns cannot be addressed until after the course is over. Situations such as these, where instructors have little or no opportunity to address concerns lead on one hand to fewer responses and greater dissatisfaction on the other (Abbott et al., 1990; Abrami et al., 2007).

The stated purpose also has an effect on what students write. In situations where students are told that the instructor would be using the information for improvements, participation and quality of responses increased while similar groups who were told that the surveys were used for promotion and tenure, muted their criticism, yet indicated greater satisfaction possibly indicating that students are far more candid about issues when the instructor is actively interested in both their feedback and in improvement, while more vague (albeit positive) when the stakes are higher (Aleamoni & Hexner, 1980; Aleamoni, 1987).

Encouraging students to complete course surveys. Faculty have a number of reservations about the motives of the students who speak out on evaluations—they either tend to be irritated by something specific, or perhaps they really like a particular instructor, but the majority do not respond because they simply have no feelings one way or the other (Feldman , 2007; Pan, et al., 2009).

Research in the field of social media and video games may provide ways of looking at how rewards, either intrinsically or extrinsically, can be used to encourage participation (Pink, 2009). Students may not participate because many do not feel that there is any feedback or response to their comments. What students want is to feel like they have contributed and that their contribution has been heard or valued. This is not to say that students need to be offered ice cream or discounts for filling out their surveys.

Social media may have the key for how to encourage students to actively participate in course evaluations. Social media and game theory look at issues of what motivates people (Pink, 2009; McGonigal, 2011). It is interesting that a student who may be asked to write 5,000 words for their class and must complete a variety of exercises of varying complexity, balks at the prospect of turning in a course survey. The process is fairly simple, and there is very little effort needed in relation to the other tasks at hand. This may be the primary problem, however. It is too simple to be perceived as little consequence, yet as Aleamoni and Hexner (1980) show, when the survey is framed as important for the instructor's career, then participation rises. Also, when students perceive that the responses may be actionable by the instructor in the

current semester and have an immediate effect on them, then instances of participation also rise (Abrami et al., 2007). What is missing is that the emphasis is generally placed on the instructor, and not the program or institution (or even the management) much of which may also influence student's thoughts and attitudes toward the particular course. For example, is it the course that a student is critical of, or the fact that they were forced into the particular class by their advisor to fulfill some financial aid requirement?

Diversifying data collection. Open-ended questions are helpful in this regard because there are few limits to frame the feedback. Instead, the feedback is factored in the background and organized with hundreds or thousands of other students who may have similar experiences, thus shedding light on a problem that may not be illustrated by more traditional methods of feedback.

An ongoing collection of feedback would change the perception of the class survey from an addendum tacked onto the end of a course and largely inconsequential, to an integrated part of campus life. Feedback in an open and ongoing system would allow for multiple methods of feedback, not just the in-class survey (See Figure 11).

A final, more draconian solution would be to make feedback an institutional requirement and set up policy where grades are not released for a class until survey forms are completed. This, however, has the possibility of negatively charging the type of feedback because of the negative and punitive tone that is set.

Another aspect of participation that can be seen in social media, is that people are much more willing to participate in a process where they can see results or at least see their own participation. Comments are a good example of this. While it may be impractical at this point to suggest that student comments be made public, some colleges have experimented with this with some success (Coleman & McKeachie, 1981). Students expressed higher satisfaction with courses that they selected where they had access to comments made by other students.

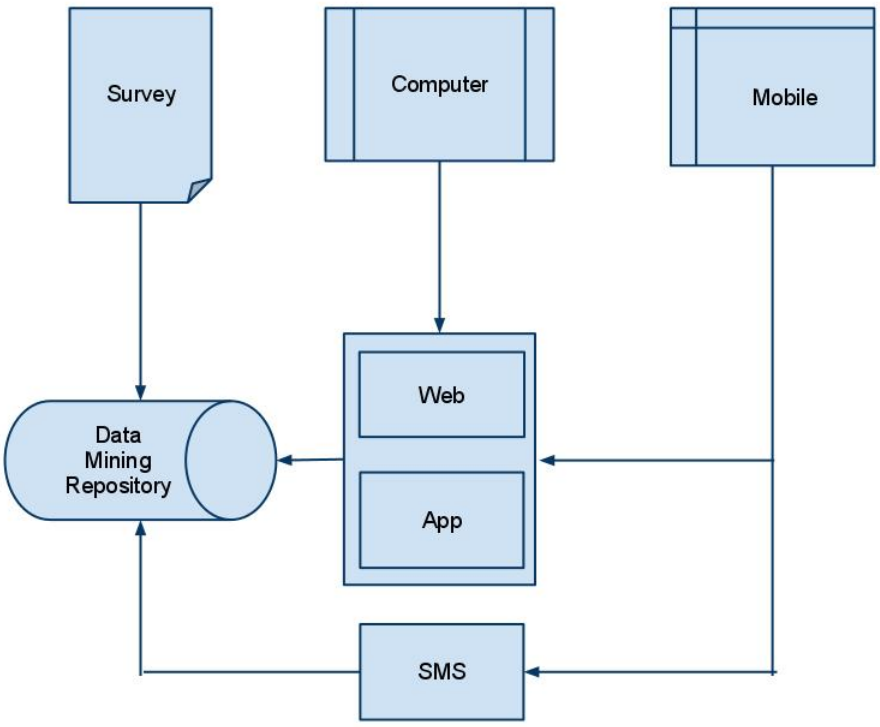

Figure 11. Text mining repository-ongoing collection model.

Another questions is, should student surveys be anonymous or not. One perspective is that if the survey is private, then comments may be anonymous; but if the comments are public, then there should be a burden of responsibility on the authors to stand up for what they are saying or endorsing. Therefore, colleges may want to consider methods of feedback that are not anonymous.

Some benefits to surveys that are not anonymous include new methods to determine the success or failure of programs in student retention. Future studies in such an environment using text mining methods could reveal through sentiment analysis thoughts and feelings that may indicate a student's dissatisfaction and thereby allow student services to intervene with more effective programs. Additionally, if comments are tracked to a student, there are additional ways that student progress can be tracked. Spelling, complexity of thought and writing, as well as a host of other issues may shed additional light onto the effectiveness of the college. Also, feedback could be encouraged in a variety of other ways that are illustrated by other types of social media. “Friending” is one such method that would allow the students to “follow” particular instructors, administrators, and even other students. Privacy, in such a system, would be a very real concern that would have to be dealt with through a combination of system restrictions and permissions as well as through institutional policies. One danger, of course, is that it may seem like a popularity contest, but students (and society in general) are becoming more used to taking control of their environment and this is one way that students can organize their own world with the

individuals that are influential to them. In such a system, social maps could then be used to find key areas of interest, influencers, and even effective social networks.

Using the data. Using the data and the policies governing its use is by far the most sensitive of the issues involved in text mining. It is made even more difficult by the fact that a more efficient and effective measure of data can also be very dangerous if used poorly, misinterpreted, or otherwise abused. While there are some colleges that have very good relationships between staff, faculty, and administration, many places suffer from lack of trust and misunderstanding that could be very destructive if the data is used the wrong way. Ideally, the data would be used to guide the institution, to create effective development of faculty and staff, and illuminate areas of optimization, new programs, or even the elimination of ineffective programs. Despite the fact that most written comments are neither lengthy, nor particularly detailed, the collection as a whole is quite remarkable in its richness.

One example of this richness can be found in Table 7 which counts the positive and negative connotations by primary and secondary content areas of the student written comment. This data, if used on the institutional level, provides an insight into what students find important. Codes that are always negative, such as instructor absences (*N*=5), equipment (*N*=20), and financial aid (*N*=2) point to items that are rarely commented on except when something has gone wrong. On the other side, expressions of thanks (*N*=39) and general comments about course design (*N*=66) are the only categories that have only positive comments and no negative comments. The top positive connotations in the code database (tables 20) give another view of

what students find important and frustrating. The code "Objectives, organization, and information" account for the most number of positive connotations, but is also the sixth highest on the table of negative connotations (see table 20).

Table 20

Top Positive Connotation by Secondary Content Areas of Student Comments

Secondary content area	Positive Connotations	% of Total Positive
Objectives, Organization, and Information	226	13.5
Instruction & Presentation	217	13.0
General	176	10.6
Knowledge & Experience	156	9.4
Assignments	148	8.9

[a] N=2110

In a general sense, this could be read to say that students value clear organization and objectives, while disliking the lack of such attributes. Indeed, the students' own words bear this out, from simply "Course was well organized" (DocID, 64), to "explaining what to expect" (DocID, 1772) and "Goal; objectives were well defined. The instructor had excellent materials" (DocID, 74) while negative comments ranged from simple "Some stated goals not addressed" (DocID, 35) to more elevated expressions of frustration with the design, "Emphasis seems to be only one what must be known to pass a test not to master the knowledge" (DocID, 34).

Table 21

Top Negative Connotation by Secondary Content Areas of Student Comments

Secondary content area[a]	Negative Connotations	% of Total Negative
Accessibility & Comfort	153	28.7
Time	62	11.6
Materials	37	6.9
Suggestion	35	6.6
Assignments	29	5.4
Objectives, Organization, and Information	27	5.1

[a] $N = 451$

Viewed in this way, the text mining procedures provide a way of looking at the metadata of student comments in a way that brings more substantive and focused comments to the top and helps develop a framework from which data can be viewed on a larger scale, independent from specific classes, instructors, and situations. Looking at the data from this level allows an administrator to get beyond specifics, and ask a more general question of: what do students want, or think that they want? What are we as an institution doing well? Where are our deficits? What are the students' perceptions of education? What do they value?

Divorced from the specific context of particular classes and instructors, this data provides an overview of a particular community that can be used in development

efforts with both new and senior faculty as well as giving administrators a high level overview of what is actually going on in their institutions.

Training and Development. Even without the above implementation, training is essential for administrators in the gathering and application of course evaluation data. Unscrupulous or erroneous uses serve only to undermine its legitimacy and effectiveness. Administration development should address:

1. Administration must understand the nature of the data that is generated from the course survey and what it can and cannot be used to do.
2. Policy must be developed around issues of collection, privacy, and access to the data.
3. The scope of the collection must be considered. Is this for just course evaluations, or can student services, library, and other segments of the college be included?
4. Uses of the data: often course evaluations are used primarily to gauge performance. Leadership will be required to use such data for development and better student satisfaction.
5. Efforts to increase student satisfaction must be tempered by the commitment to academic excellence.
6. Change needs to be managed. Collection of data and the redesign of this process must have input from all levels: administration, faculty and staff.

Suggestions for Further Research

This study represents an initial step. Text mining offers a rich area for exploration of unstructured data that was previously out of reach for systematic institutional analysis beyond an occasional qualitative study. While approaching student written comments using data and text mining tools offers a unique look into what students value about their educational experiences, levels of engagement, and other contextual aspects relating to the institution, there are new questions about alternative methods of data collection and whether or not these alternatives could produce better results though more timely observations or more flexible and varied forms of input.

Other productive directions could be aimed at how such data may be used constructively by institutions. The information presented in this research could form the basis of a model for faculty development and institutional development. As text mining tools and algorithms become more sophisticated and begin to "learn" through multiple iterations and benchmarking (see Figure 10) the information that can be teased out of the data will require new models of analysis as well as new policies to control the use and dissemination of that data. Academic analytics to provide evidence of learning, identifying patterns of behavior, historical analysis, forecasting, and risk analytics are all areas of concern in this area.

Summary and Conclusion

Course evaluations are only one component to the data generated by a college. It is important to remember that it is a collection of perceptions and sentiment that ranges from astute observation to uncritical monosyllabic responses. Despite that, when assessed en mass, through processes like text mining, a larger picture emerges. As colleges struggle to be more responsive to their students, this data becomes more important. A reimagining of the feedback system from the current, pencil-and-paper system into a real-time collection of student feedback could be a monumental shift in how colleges respond to student concerns and are able to quickly adjust to a variety of issues, not only academic, but also touching on other areas of college life. Individually, student written comments may not have a lot to say regarding some of the broader issues facing colleges, but examined in their totality, there are larger patterns that can, if properly read, be used to guide the college and its own internal development, forecasting trends, and conducting risk analytics.

REFERENCES

REFERENCES

Abbott, R., Wulff, D., Nyquist, J., Ropp, V., & Hess, C. (1990). Satisfaction with processes of collecting student opinions about instruction: The student perspective. Journal of Educational Psychology, *82*(2), 201-206. http://search.ebscohost.com.ezproxy.lib.csustan.edu:2048.

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), The scholarship of teaching and learning in higher education: An evidence-based perspective (pp. 385–456). New York: Springer.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. Instructional Science. *9*(1), 67-84.

Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. In L. M. Aleamoni (Ed.), Techniques for evaluation and improving instruction, new directions for teaching and learning (No. 31, pp. 25–31). San Francisco: Jossey-Bass.

Alhija, F. N. A., & Fresko, B. (2009). Student evaluation of instruction: What can be learned from students' written comments? Studies in Educational Evaluation. *35*(1), 37-44.

Angus, J. (2003). Data mining engine revs up. Infoworld, *25*(45), 44-45. Retrieved from EBSCOhost.

Bishop, C. (1995). Neural networks for pattern recognition. Oxford, UK: Oxford University Press.

Braskamp, L. A., Caulley, D. & Costin, F. (1979). Student ratings and instructor self-ratings and their relationship to student achievement. American Educational Research Journal, *16,* 295–306.

Braskamp, L., Ory, J., & Pieper, D. (1981). Student written comments: Dimensions of instructional quality. Journal of Educational Psychology, *73*(1), 65-70. http://search.ebscohost.com.ezproxy.lib.csustan.edu:2048.

Campbell, J., De Blois, P. B., & Oblinger, D. (2007). Academic analytics: A new tool for a new era. Educause Review, *42*(4), 42-57.

Centra, J. A. (1987). Formative and summative evaluation: Parody or paradox? In L. M. Aleamoni (Ed.), Techniques for evaluation and improving instruction, new directions for teaching and learning (No. 31, pp. 47–55). San Francisco: Jossey-Bass.

Chen, C. M., & Chen, M. C. (2009). Mobile formative assessment tool based on data mining techniques for supporting web-based learning. Computers & Education. *52*(1), 256-273.

Coffman, W. (1954). Determining students' concepts of effective teaching from their ratings of instructors. Journal of Educational Psychology, *45*(5), 277-286. http://search.ebscohost.com.ezproxy.lib.csustan.edu:2048.

Coleman, J., & McKeachie, W. J. (1981). Effects of instructor/course evaluations on student course selection. Journal of Educational Psychology. *73*(2), 224-26.

Creswell, J. W. (2003). Research design: Qualitative, quantitative, and mixed method approaches. Thousand Oaks, CA: Sage Publications.

Downey, J. (2003, September). Emotional awareness as a mediator of community college student satisfaction ratings. Community College Journal of Research & Practice, *27*(8), 711. Retrieved May 11, 2009, from Academic Search Elite database.

Feldman, K. A. (1976)a. Grades and college students' evaluations of their courses and teachers. Research in Higher Education, *4*, 69–111.

Feldman, K. A. (1976)b. The superior college teacher from the students' view. Research in Higher Education. *5*(3), 243-88.

Feldman, K. A. (1989b). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multi-section validity studies. Research in Higher Education *30*(6): 583–645.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), The scholarship of teaching and learning in higher education: An evidence-based perspective (pp. 93–143). New York: Springer.

Feldman, M. J. (1993). Factors associated with one-year retention in a community college. Research in Higher Education. *34*(4), 503-12. Retrieved May 5, 2009,

from http://www.springerlink.com.ezproxy.lib.csustan.edu:2048/content/u3524v 523080k83u/

Flick, Uwe. (2006). An introduction to qualitative research. London: Sage Publications.

Franklin, U., & M. Theall (1989). Rating the readers: Knowledge, attitude, and practice of users of student ratings of instruction. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Fresko, B., & Nasser, F. (2001). Interpreting student ratings: Consultation, instructional modification, and attitude towards course evaluation. Studies in Educational Evaluation. *27*(4), 291-305.

Gallagher, T. (2000). Embracing student evaluations of teaching: A case study. Teaching Sociology, *28*(2), 140–147.

Glaser, B., & Strauss, A. (1967). Discovery of grounded theory. Strategies for Qualitative Research. London: Sociology Press.

Guthrie, E. R. (1954). The evaluation of teaching: A progress report. Seattle: University of Washington.

Hardy, N. (2003). Online ratings: Fact and fiction. New Directions for Teaching and Learning. *69*, 31-38.

Hodges, L. C., & Stanton, K. (2007). Translating comments on student evaluations into the language of learning. Innovative Higher Education. *31*(5), 279-286.

Johnson, T.D. (2003). Online student ratings: Will students respond?, New Directions for Teaching and Learning *96*, pp. 49–59.

Kember, D., & Wong, A. (2000). Implications for evaluation from a study of students' perceptions of good and poor teaching. Higher Education, *40*(1), 69–97.

Kilpatrick, W. H. (1918). The project method. Teachers College Record, *19*(4), 319-335.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics *33*:159-174.

Lewis, K. G. (2001), Making sense of student written comments. New Directions for Teaching and Learning, 2001: 25–32.

Lim, J., Kim, M., Chen, S., & Ryder, C. (2008, June). An empirical investigation of student achievement and satisfaction in different learning environments. Journal of Instructional Psychology, *35*(2), 113-119.

Lin, Y., McKeachie, W. J., & Tucker, D. G. (1984). The use of student ratings in promotion decisions. Journal of Higher Education, *55*, 583–589.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damerau (Eds.), Handbook of natural language processing. (pp. 627-666). Boca Raton, FL: Chapman & Hall/CRC.

Luan, J., Zhao, C. M., & Hayek, J. C. (2009). Using a data mining approach to develop a student engagement-based institutional typology. IR Applications, Volume 18, February 8, 2009. Association for Institutional Research.

Manochehri, N., & Young, J. (2006, Fall 2006). The impact of student learning styles with web-based learning or instructor based learning on student knowledge and satisfaction. Quarterly Review of Distance Education, *7*(3), 313-316.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. British Journal of Educational Psychology, *52*(1), 77–95.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students for different academic settings and their relationship to student/course/instructor characteristics. Journal of Educational Psychology, *75*(1), 150–166.

Marsh, H. W. (1984). Students' evaluations of university teaching; Dimensionality, reliability, validity, potential biases and utility. Journal of Educational Psychology, *76*, 707–754.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, *11*, 253–388.

Marsh, H. W. (1989). Responses to reviews of students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. Instructional Evaluation *10*: 5–9.

Marton, F., Dall'Alba, G., & Betty, E. (1993). Conceptions of learning. International Journal of Educational research. *19*(3), 277-300.

McGonigal, J. (2011). Reality is broken: Why games make us better and how they can change the world. New York: Penguin Press.

Nisbet, R., Elder, J. F., & Miner, G. (2009). Handbook of statistical analysis and data mining applications. Amsterdam: Academic Press/Elsevier.

Ory, J. C. (2000). Teaching evaluation: Past, present, and future teaching, new directions for teaching & learning (No. 83, pp. 13–18). San Francisco: Jossey-Bass.

Pan, D., Tan, G., Ragupathi, K., Booluck, K., Roop, R., & Ip, Y. (2009). Profiling teacher/teaching using descriptors derived from qualitative feedback: Formative and summative applications. Research in Higher Education, *50*(1), 73-100. Retrieved July 23, 2009, doi:10.1007/s11162-008-9109-4

Pink, D. H. (2009). Drive: The surprising truth about what motivates us. New York, NY: Riverhead Books.

Reisenzein, R. (1983). The Schachter theory of emotion: two decades later. Psychological Bulletin, *94*, 239-64.

Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction, Research in Higher Education *46*(8), pp. 929–953.

Romero C., & Ventura S. (2007). Educational data mining: A survey from 1995 to 2005, Expert Systems with Applications *33*(1), pp. 135–146.

Saljo, R. (1979). Learning in the learners, I—Some commonsense conceptions. Reports from the Institute of Education. University of Gothenburg, No. 77.

Salovey, P., Mayer, J. D., Goldman, S. L., Turvey, C., & Palfai, T. P. (1995). Emotional attention, clarity, and repair: Exploring emotional intelligence using the trait meta-mood scale. In J. Pennebaker (Ed.), Emotion, disclosure, and health. Washington, DC: American Psychological Association.

Schachter, S., & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. Psychological Review, *69*(5), 379 - 399.

Sheehan, E., & DuPrey, T. (1999). Student evaluations of university teaching. Journal of Instructional Psychology, *26*(3), 188. http://search.ebscohost.com.ezproxy.lib.csustan.edu:2048

Spicer, S. L. (1989). Paths to Success. Volume one: Steps towards refining standards and placement in the English curriculum. Glendale, CA: Glendale Community College. (ERIC Document Reproduction Service No. ED 312 021).

StatSoft, Inc. (2008). STATISTICA (data analysis software system), version 8.0 www.statsoft.com.

Theall, M., & Franklin, J. (Eds.). (1990). Student ratings of instruction: issues for improving practice, new directions for teaching and learning (No. 43). San Francisco: Jossey-Bass.

Theall, M., & Franklin, J. (1991). Using student ratings for teaching improvement. In M. Theall & J. Franklin (Eds.), New Directions for Teaching and Learning, *48*, (pp. 83-96). San Francisco: Jossey-Bass.

Theall, M., & Franklin, J. (1999). Faculty thinking about the design and evaluation of instruction. In P. Goodyear & N. Hativa (Eds.), Teacher thinking, beliefs, and knowledge in higher education. The Netherlands: Kluwer Academic Publishers.

Yu, C. H., DiGangi, S. A., Jannasch-Pennell, A., Lo, W., & Kaprolet, C. (2007). A data-mining approach to differentiate predictors of retention. In the Proceedings of the Educause Southwest Conference, Austin, Texas, USA.

Zhao, C. M., & Luan, J. (2006). Data mining: Going beyond traditional statistics. New Directions for Institutional Research. *131*, 7-16.

Zimmaro, D. M., Gaede, C. S., Heikes, E. J., Shim, M. P., & Lewis, K. G. (2006). A study of students' written course evaluation comments at a public university. http://www.utexas.edu/academic/mec/publication/pdf/fulltext/courseevalcomments.pdf

APPENDICES

APPENDIX A

CODEBOOK DESCRIPTIONS WITH EXAMPLES

Code	Description	Positive Examples	Negative Examples
Absent	Comments that refers to either instructor absence, leaving early, or otherwise being unavailable.	N/A	Being let out every night before 8pm or just at 8pm. I felt cheated. (DocID, 2248)
Clarity	Comments referencing either confusion or clarity of various aspects of the course.	The instructor was very knowledgeable about everything and explained it very well (DocID, 169)	Instructions for assignments were unclear - verbally confusing waht was expected /online posting of who was missing assignments unnecessary. Contact student directly, not publicly (DocID,227)
Competence & Professionalism	References to the competence and professionalism of the instructor	The instructor was an excellent teacher whom was very well prepared for class (DocID, 72)	Instructor needs more work experience in teaching, instructing classroom whatever you want to call it. (DocID,131)
Expression of thanks	Comments that indicate an appreciation or expression of	Thank you. This class has opened my eyes.	N/A

	thanks	(DocID, 2309)	
General	References to the instructor in general, but not applicable to any of the other categories.	The instructor was great at her job and would like to have her again (DocID, 255)	Need the right teachers for the right class. (DocID, 2389)
Helpful, Timely, Feedback & Flexible	References to the instructor as either helpful or flexible in their dealings with students, both in and out of class. Timeliness refers to references to how quickly the instructor responds to the student or returns work that the student has turned in.	Instructor was excellent; patient, understanding, compassionate. (DocID, 6)	It seems that many questions were not heard or responded to. (DocID, 150)
Instruction & Presentation	Specific references to the presentation and instruction of the instructor or in-class instruction. Also, general comments such as "good" or "Great" for question 11 dealing with instructor would be coded in this area.	Very well presented (DocID, 114)	I would have liked more classroom instruction (DocID, 566)
Knowledge & Experience	Specific references to instructor's knowledge and experience or being informative.	The instructor was an excellent teacher who was always prepared and very knowledgeable about her subject. (DocID, 36)	Teacher lack of Knowledge (DocID, 2010)
Multiple Instructors or Guests	Used when student indicates that there was either more than one instructor, or if there were guest speakers.	I really appreciated the person who came into our class to speak about HIV (DocID,1843)	It was difficult having three different teachers for this course (DocID, 59)

Organization and Focus	Specific references to the instructor (as opposed to the course) as being focused, on task, or organized, methodical, or well-paced. Also, may include references to classroom management and being prepared.	The strongest feature off this courses was the binder put together as a resource for students (DocID, 1595)	Class was disorganized, teacher knew her subject, but changed what we were to do in class all the time (DocID, 51)
Personality	References to the personality of the instructor: being patient, understanding, interesting, enthusiasm (as well as a lack of these traits)	The instructor is smart, very funny, and has excellent advice. (DocID, 32)	She had a bad ass case of the I's and I this, I that, My this, my that. Can't rate her (DocID, 295)
Rigor	References to the instructor's rigor through both positive and negative references to toughness, pushing, amount of homework, grading, and setting standards.	This was the 1st real instructor we have had who held us accountable for learning and completing work. This made the less committed students unhappy. (DocID, 165)	No Homework (DocID, 1912)
Assignments	References to assignments, homework, or tests. A positive reference to homework is coded as CDASG while a reference to high standards, pushing or effective use of tension (as well as a lack of) in homework by an instructor is coded as IINSRIG	Helpful to have assignments due on a weekend- good balance between written and online discussion (DocID, 825)	Maybe a quiz or test at the end of class could have been upgraded just to test how well we understood the information (DocID, 26)
General	Vague and general references, either positive or negative, such as "good", "Great", or "Excellent" in questions that	It was a great class. I really enjoyed it. (DocID, 2375)	N/A

	deal specifically with the course (as opposed to the instructor) or as responses to general open-ended questions dealing with the course.		
Materials	References to specific course materials or resources such as textbooks and binders or handouts with additional tools or information that is not specifically referenced as an assignment.	Enjoyed the material that the instructor brought to class (DocID, 174)	didn't really use the text (DocID, 47)
Objectives, Organization, and Information	References about information, organization, and course objectives or outcomes that that do not specifically reference instruction or the instructor. Instead, the overall nature of the course design is implied or expressed.	It exceeded my expectations, provided a great deal of "hands-on" information (DocID, 972)	Emphasis seems to be only one what must be known to pass a test not to master the knowledge (DocID, 34)
Suggestion	Suggestions for improving the course or the student experience by adding or removing assignments, features, materials, books, etc. (with the exception of time)Suggestions may also include veiled criticism such as "not many handouts, we need more of them."	N/A	It would have been helpful to do the work groups spread out throughout the course instead of all at once at the end (DocID, 1019)
Time	A specific subset of suggestion, these are comments that specifically reference the need for more or	Excellent instruction....class should have been longer unable to cover the	Not enough time to fully grasp assessment and treatment plans to

	less time in a variety of aspects of the course.	material in necessary depth. (DocID, 229)	feel totally confident is getting a doing a job in this field (DocID, 965)
Accessibility & Comfort	References to either the comfort or discomfort of a room or the environment, accessibility (such as ADA issues), or other environmental issues. Often, issues of comfort are implied accessibility issues.	142 nice big room (DocID, 430)	From bad to worse - stairs everywhere - handicapped people including I w/o wheelchair and no elevator. (DocID, 344)
Equipment	References to equipment in rooms or labs such as computers, overhead projectors, chairs, desks, or electrical outlets as well as other forms of technology like online forums or LMS.	Computer.[In context of the question, "what is the strongest feature of the class?"] (DocID, 1462)	Only problem was the screen that blocked the chalkboard (DocID, 442)
Financial Aid	Specific references to financial aid	N/A	except for my loan process with […], I believe it was due to my own inexperience and lack of information. (DocID, 2628)
General	General references in response to Question 20 about the institution as well as comments dealing with the campus and other interactions with the institution not specifically related to the course or instruction.	UOP [the university] has been great (DocID, 218)	This is a very confusing campus. I'm glad most of the classes are in the same building. (DocID, 338)

Registration	Specific references to the registration process or support staff relating to the registration process. Most general responses dealing with Question 5	Since I was applying from a distance, everyone made a real effort to explain things and be flexible. (DocID, 1)	Sometimes office workers are dismissive and officious (DocID, 323)
Support	References to support staff other than registration or instruction.	The staff is very accommodating and helpful (DocID, 327)	Wrong information an some things given out at the office (DocID, 460)
Interaction with other students	This refers to comments focusing on working with other students, other student's behavior, or presenting group or individual projects where they are observed or get feedback from peers		
Reference to Future	Comments made by students that reference some aspect of the future, either in future courses or their future careers or personal lives.	The information attained will be helpful to me in the future. (DocID, 520)	I like the accelerated classes but also worry that I don't have enough knowledge of the subject (DocID, 2148)
Self Referential	Any excerpt that refers to the writer-either self reflection, self acknowledgement, observations about how the class, material, or instruction impacts them in a personal way.	Yes, I have a history of counseling theory knowledge, but it built upon what I already knew (DocID, 997)	I do not feel confident that I could properly assess or treatment plan someone based on the speed in which we went through this subject (DocID, 53)

APPENDIX B

CUSTOM CODING DATABASE

The references to specific coded statements from the written comments on the courses surveys within this dissertation directly relate to a specific ID number assigned to each statement during the coding process. Each statement was assigned a DocID number from 1 to 2661. The primary function of this ID number was to maintain the relationship between the statements, the codes assigned to those statements, and the descriptor data for each of the surveys from which those statements come. Because each one is unique and refers to only one specific statement, it also serves as a useful identifier when referencing a particular statement within this dissertation. Therefore, each statement can be cited from the coding database using that unique identifier. For example, "It was all strong. I think what hit home for me was having a personal story shared. I learned a lot" (DocID, 1490).

The original intent for coding the student written comments was to use a program such as HyperResearch or Dedoose. After some comparisons, the initial coding was done in the web-based program Dedoose (http://www.dedoose.com). The data had to be reformatted from the original Excel files so that the descriptors remained in Excel, and all of the written comments were entered on separate Word documents. After encountering problems with linking the descriptors with the data, the researcher submitted a support ticket to the company requesting a way to upload

the data through a batch process. The batch process allowed for the mass uploading of the Word documents, but the link between the descriptors and data took, on average, 90 seconds for each link and another 2-3 minutes for each code roughly equating to 200 hours of coding and linking. In an email from the lead developer of Dedoose, he stated:

> It's apparent your[sic] coming from the quantitative world and integrating the qualitative, where most of our users and our workflow is designed for qualitative researchers moving to integrate quantitative, which is why I feel your experiencing a serious workflow impediment. Where your working with a ton of very small documents, most of our qualitative researchers work with a handful (30-200) of large documents (30k-80k characters). So for you, since there are so many documents with very little content in them it creates an incredible burden to link them all together. (J. Taylor, personal communication, September 23, 2010)

Based on this assessment, the researcher decided to create a coding program using Microsoft Access which allowed for a much faster coding process and greater control over the reports. Two tables, named Descriptors and Documents, were created with a similar structure to the original Excel spreadsheet used during the transcription process. The Codes table was created to store coding information and the Link table was created in order to allow for multiple codes to be assigned to the same phrase.

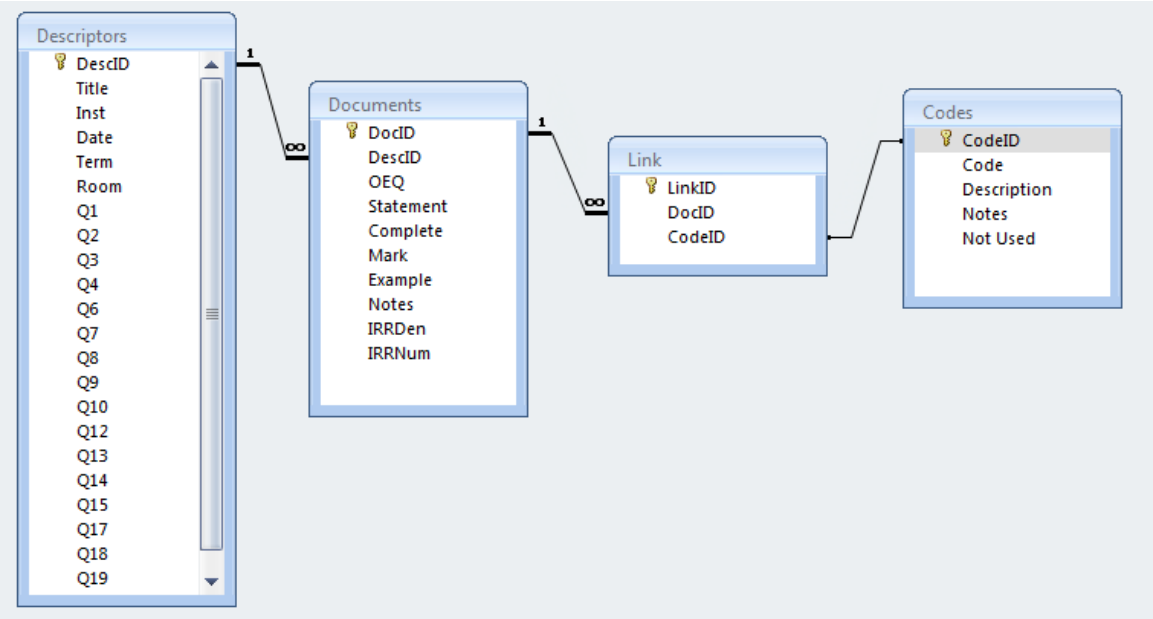

Figure 12. Tables and relationships in coding database.

Several queries were written in order to frame three primary forms used in the coding process: The main coding page (see Figure 13) a summary form used to quickly find statements by code (see Figure 14), and an integrator reliability form used in reconciling codes from multiple readers (see Figure 15).

The main coding page allows the user to cycle through each response and either 1) split the response into multiple meaning units, 2) quick apply a positive or negative code, 3) apply other codes from the codebook, 4) add a new code to a codebook, 5) add a note regarding coding decisions, or 6) apply filters to view all comments, completed comments with codes, or all comments without codes.

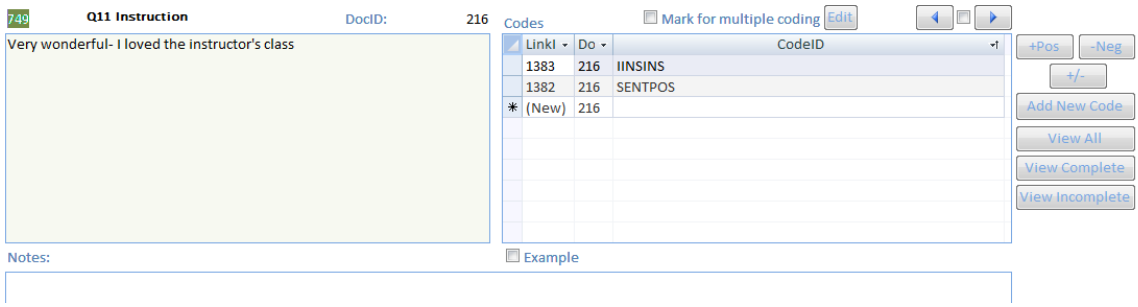

Figure 13. Main coding page

The coding summary page was created in order to organize each statement by code. This page is primarily used to drill down to specific statements within a code group, however, it was also used to help define the code as well using the notes feature.

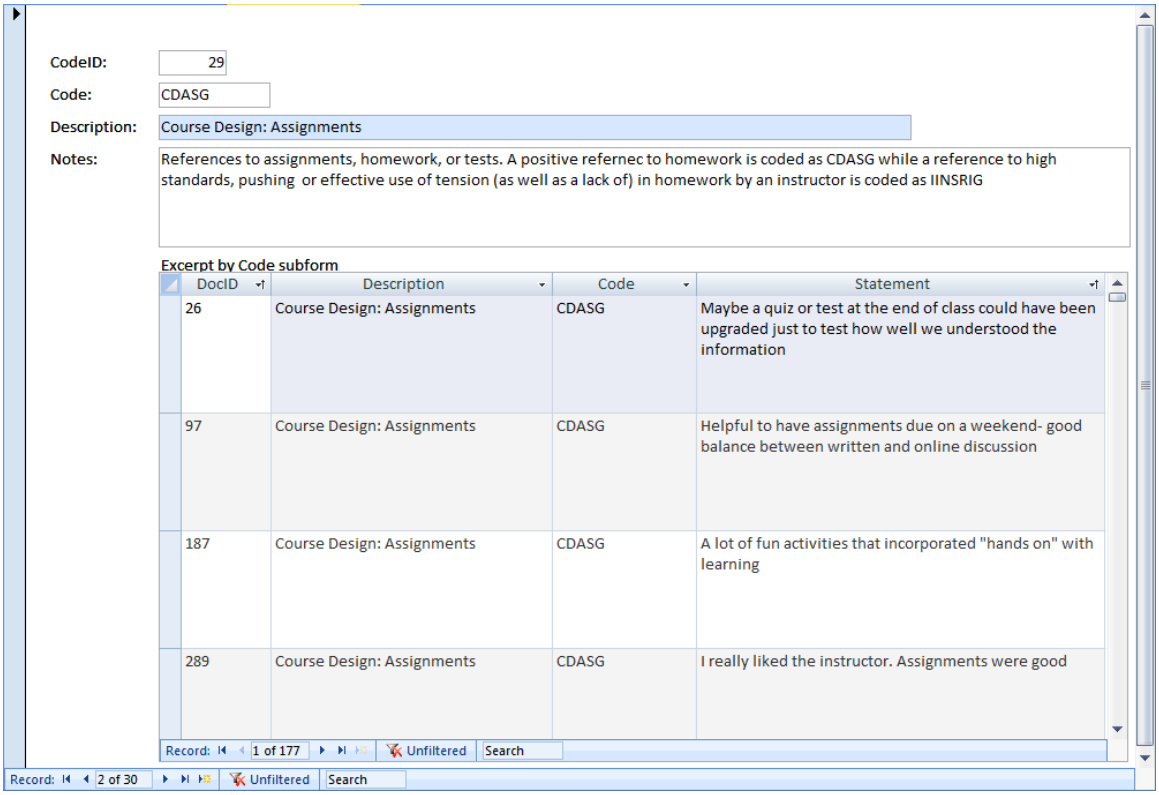

Figure 14. Coding summary form.

An interrater reliability form was also created that automatically selected 100 random statements and presented them in the same way as the original coding was done (see Figure 13). Once the second rater completed his coding, the interrater reliability form was used to quickly compare the original coding with that of the second rater. In cases of multiple raters, this process could be used multiple times with either the same dataset or with a new random selection. Once completed, results were exported to SPSS for the interrater reliability statistics that were reported in chapter 4.

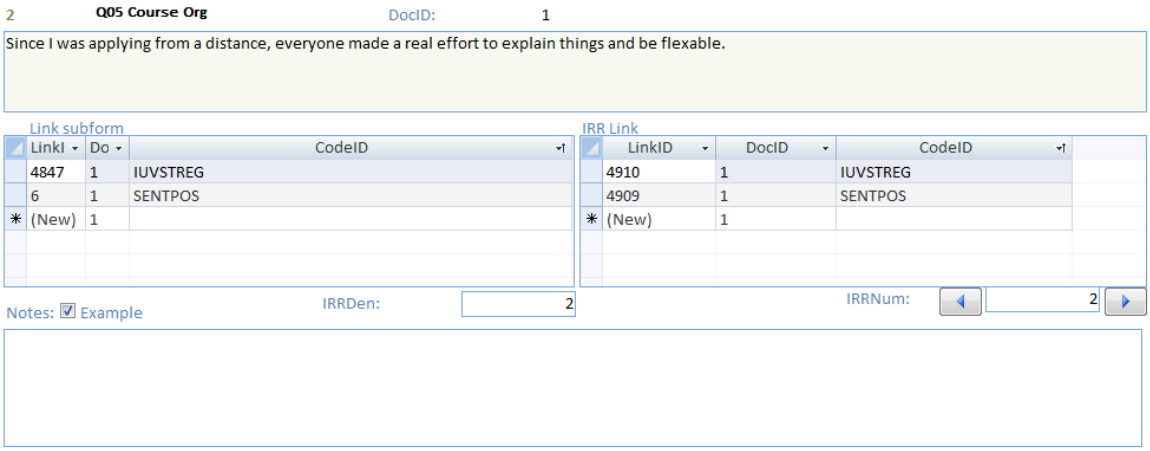

Figure 15. Interrater reliability form.

Finally, the following export query was written in order to export the completed data from the tables to the format required by both SPSS for analysis:

SELECT Descriptors.DescID, Descriptors.Title, Descriptors.Inst, Descriptors.Date, Descriptors.Term, Descriptors.Room, Descriptors.Q1, Descriptors.Q2, Descriptors.Q3, Descriptors.Q4, Descriptors.Q6, Descriptors.Q7,

Descriptors.Q8, Descriptors.Q9, Descriptors.Q10, Descriptors.Q12, Descriptors.Q13, Descriptors.Q14, Descriptors.Q15, Descriptors.Q17, Descriptors.Q18, Descriptors.Q19, Descriptors.Complete, [qryStatements by code and Descriptor_Crosstab].[Total Of Statement], [qryStatements by code and Descriptor_Crosstab].CD, [qryStatements by code and Descriptor_Crosstab].CDASG, [qryStatements by code and Descriptor_Crosstab].CDMAT, [qryStatements by code and Descriptor_Crosstab].CDSUG, [qryStatements by code and Descriptor_Crosstab].CDTIME, [qryStatements by code and Descriptor_Crosstab].CDUND, [qryStatements by code and Descriptor_Crosstab].FUTR, [qryStatements by code and Descriptor_Crosstab].IENVACCOM, [qryStatements by code and Descriptor_Crosstab].IENVEQ, [qryStatements by code and Descriptor_Crosstab].IINSABS, [qryStatements by code and Descriptor_Crosstab].IINSCLA, [qryStatements by code and Descriptor_Crosstab].IINSCOM, [qryStatements by code and Descriptor_Crosstab].IINSGEN, [qryStatements by code and Descriptor_Crosstab].IINSHLP, [qryStatements by code and Descriptor_Crosstab].IINSINS, [qryStatements by code and Descriptor_Crosstab].IINSKNO, [qryStatements by code and Descriptor_Crosstab].IINSMUL, [qryStatements by code and Descriptor_Crosstab].IINSORG, [qryStatements by code and

Descriptor_Crosstab].IINSPER, [qryStatements by code and Descriptor_Crosstab].IINSRIG, [qryStatements by code and Descriptor_Crosstab].IINSTHK, [qryStatements by code and Descriptor_Crosstab].ISTU, [qryStatements by code and Descriptor_Crosstab].IUVST, [qryStatements by code and Descriptor_Crosstab].IUVSTFA, [qryStatements by code and Descriptor_Crosstab].IUVSTREG, [qryStatements by code and Descriptor_Crosstab].IUVSTSUP, [qryStatements by code and Descriptor_Crosstab].SELFCRIT, [qryStatements by code and Descriptor_Crosstab].SENTNEG, [qryStatements by code and Descriptor_Crosstab].SENTPOS FROM Descriptors LEFT JOIN [qryStatements by code and Descriptor_Crosstab] ON Descriptors.DescID = [qryStatements by code and Descriptor_Crosstab].DescID;"