Text Analysis and Word Frequency Examination of Student Evaluations: Informing the Relationship between Language Sentiment, Instructor and Course Traits

Kyle Breznicker
Ryan Gomez

This honors paper is submitted to the University of Oregon Department of Economics, in partial fulfillment of the requirements for honors in Economics

14 June 2019

Under the supervision of Bill Harbaugh
University of Oregon

Abstract

This study utilizes text responses gathered from student evaluations from the University of Oregon, to provide descriptive statistics and analysis about the type and nature of language used to give feedback to university faculty. Our goal was to determine if that language was biased, and to provide insight on the value of text evaluations within student responses. Using two text analysis programs, LIWC and WordStat, along with regression analysis, we aggregate metrics for positive and negative sentiment to examine relevant trends in our data set and expand on the findings of previous papers. This paper contains explanations of both programs' processes. We find no evidence of gender bias in either word frequency use or sentiment differences. We find mixed results in instructor rank and course subject differences. We draw conclusions based on most frequently used words and our regression output to find that the text from student evaluations has strong potential to be an improved source for institutions to make decisions regarding instructor performance compared to potentially biased numerical score systems.

Table of Contents

Introduction:

Academic institutions attempt to quantify instructor and class quality through various information gathering techniques. These institutions hope to create a metric with which to judge and ultimately rank instructors for their own purposes. Whether they are considering professors for tenure or possible promotions or identifying valuable courses, institutions need to have a consistent way of measuring teaching staff against one another. Many institutions, including the University of Oregon, utilize student evaluations to fill this role. These evaluations, commonly referred to as SETs, give students the opportunity to rate the courses they take. Recently, however, the deficiencies and potential bias of SETs has become a focal point in the push to reform instructor evaluations across academia. Researchers have identified that using only numerical based evaluations provides a flawed snapshot of instructor performance.

In 2017, the University Senate sought further research into the validity of the overall evaluation program. This prompted a task force with the goal of overhauling student evaluations. Evaluations at the University of Oregon have followed the same general format since the university modernized its evaluation system in 2007 (https://registrar.uoregon.edu/course-evaluations). By establishing an online portal to gather student feedback, the university created a way of making its data available for analysis. The data includes text and numerical responses, allowing for a variety of data analysis projects to be conducted. Using UO registrar and faculty data, organized by the private technology firm collegenet, this paper will process thousands of student evaluations. We aggregated the text and applied sentiment text analysis to qualify the statements. The program we will use for the text analysis is Linguistic Inquiry and Word Count, commonly referred to as LIWC. The goal of this software is to provide robust measures of text

traits for analysis. In order to avoid as much personal bias as possible, the data will be sorted using LIWC's established academic dictionaries.

Additional descriptive analysis of our data was conducted using WordStat (https://provalisresearch.com/products/content-analysis-software/), a program designed to calculate word frequencies in bodies of text. Word frequency analysis focuses on differences in how students write about male and female professors, and the various instructor ranks. We use the most frequent words in our data set to identify any gender bias present in the text evaluations.

Our primary research goal is to assess the general tone and identify any bias that is present in the surveys. We hope that using text analysis will give the University of Oregon a different perspective on their evaluations, as opposed to only using numerical based techniques. The second goal of this paper is to provide an assessment of the types of words that are being used in the student evaluations. Additionally, this paper will provide a commentary on the robustness of student evaluations in general, applicable to other academic institutions.

Literature Review:

Questions about the legitimacy and potential bias of the instructor evaluation system at the University of Oregon have been examined by a previous group of economics students. In a paper titled C*ourse Evaluations, Teaching Quality and Student Achievement (*University of Oregon 2017*),* Kenneth Ancell and Emily Wu analyzed data similar to what we will be using from the university to see if numerical SET scores are an accurate measure of teacher quality and to look for gender biases. They examined the impact of instructor characteristics like race and gender on the numerical SET scores they received, and analyzed the impact of student and class factors on how students rate instructors. Student characteristics included measures of academic

ability as well as traits like race, gender, class standing, and major. The instructor characteristic category included SET scores, gender, race, age, and whether or not they are a grad student. The class factors category included course quality evaluation answers, department, year, class average and class size. Ancell and Wu then created a regression to measure how these categories impact both the average course GPA and student achievement in similar future courses. They found that female instructors receive systematically lower course evaluations despite their students having higher degrees of success in subsequent courses.

In our evaluation, we hope to expand on the analysis of the bias found by Ancell and Wu by examining the language used by students when evaluating instructors with textual analysis software and techniques to see if there is a similar degree of gender bias present in written evaluations. Quantitative text analysis is an emerging field that has become increasingly relevant in social science, computer science, and psychology. In his 1997 paper titled "A Conceptual Framework for Quantitative Text Analysis" published in *Quality and Quantity: International Journal of Methodology*, Carl Roberts provides a general overview of the emerging role of text analysis in academic research and describes the fundamental characteristics of the various methods being used. He describes text analysis as "the application of one or more methods for drawing statistical inferences from text population (...) in which texts are interpreted either instrumentally (according to the researcher's conceptual framework) or representationally (according to the texts' sources' perspectives), as well as in which variables are thematic (counts of word/phrase occurrences), semantic (theses within a semantic grammar), or network-related (theme- or relation-positions within a conceptual network)" (Roberts 1997, 259). These methods can be used to create a matrix of relationships between text, semantics, and context to create valid quantitative measures of qualitative text traits and themes.

Roberts also lays out a timeline of modern text analysis techniques. The earliest forms of text analysis relied primarily on counting the occurrences of various content categories and creating a data matrix to estimate the relationships between themes, then developing explanations of why these relationships existed post hoc. More contemporary methods of analysis have introduced instrumental methods designed to account for the contextual factors by "[identifying] individual and societal characteristics about which society members may be unaware" to accompany representational methods "used to characterize text in ways that their source intend them to be understood" (Roberts 1997, 263). Roberts concludes that sound text analysis is based on a data matrix that accounts for both text and context related variables. This paper helped inform the broad approach and methodology to text analysis used by the creation of the LIWC software we will be using for this project and is cited as a key reference by the program's developers.

Quantitative text analysis research has been done recently on methods of evaluations that are relatively similar in general principal to the SET system used by the University of Oregon. In their paper titled *Predicting a Business' Star in Yelp from Its Reviews' Text Alone* Mingming Fan and Maryan Kahademi (ArXiv, 2014) develop and test a few different models of text analysis to create a linear regression capable of predicting the Yelp star score a restaurant would receive based on the written text of the review. They used the "bag of words" method in which they took a large body of text from reviews and generated variable for the top K (amount of) words of various types. The results showed that the linear regression that isolated the most frequently used adjectives consistently performed best at predicting scores, which makes intuitive sense given that they are the words most likely to be associated with negative or positive emotions.

Similar methods have been used to analyze sentiment and opinion from the short blocks of text that are posted on Twitter. In their paper titled *Twitter as a Corpus for Sentiment Analysis and Opinion Mining* (2010, LREC) Alexander Pak and Patrick Paroubeck collected 300,000 text posts and demonstrated a method for sentiment classification using this data, then tested that classification method on another set of tweets. In describing their methods the authors wrote "we used the collected corpus to train a sentiment classifier. Our classifier is able to determine positive, negative and neutral sentiments of documents" (Pal, 1326). They also used a variety of factors in text classification including emoticon usage, superlative adverbs, utterances and superlative adjectives. They also introduced the length of n-grams (blocks of text with n number of words in them) as a relevant variable to test how many words should be grouped together and assigned negative, positive or neutral value. They found that uni-grams (single words) and bi-grams (groups of two words) provided the most accurate results, and their adaptation of this model proved effective at measuring the sentiment of large groups of text.

A question that we will be using our data and linguistic analysis to answer is whether or not gender bias is present in the student evaluations. Anna Kaatz and her team authored a paper that tackled a very similar topic. *Analysis of NIH R01 Application Critiques, Impact and Criteria Scores: Does the Sex of the Principal Investigator Make a Difference?* (2016, Academic Medicine), asks if gender bias had an impact on the critiques of research project grant requests at her institution. It is an expansion on a Kaatz paper from a year earlier that "suggests that text analysis of grant critiques may be useful for identifying potential gender bias in peer review" (Kaatz et al 2016). The paper analyzed 739 such critiques, from requests that were ultimately funded as well as those that were not.

The method the authors used is very similar to what we are going to pursue with LIWC. The team identified several categories of words, “ability, achievement, agentic, negative evaluation, positive evaluation, research, and standout adjectives” (Kaatz et al 2016), and analyzed the differing presence of words in these categories across reviews of male and female applicants. What Kaatz found is that despite using more words categorized as positive, reviewers still scored requests authored by female applicants lower, “Critiques of female PIs’ Type 2 applications were linguistically stronger, more often containing standout adjectives and words about ability” (Kaatz et al 2016). The authors conclude this is evidence that the structure of the reviewing system will “lead reviewers to implicitly hold male and female applicants to different standards of evaluation” (Kaatz et al 2016). This paper used category based text analysis to look for gender bias in critiques of project proposals. Our paper will use this technique, based in LIWC to assess for potential bias in the University of Oregon’s student evaluation surveys.

A previous paper also written by Kaatz, *A Quantitative Linguistic Analysis of National Institutes of Health R01 Application Critiques from Investigators at One Institution* (Academic Medicine 2015), provides a list of LIWC dictionaries that will be useful for our analysis. The authors write that these categories are directly applicable to analyzing evaluations. They describe their dictionaries as follows, “These word categories are “ability” (e.g., skilled, expert, talented), “achievement” (e.g., honors, awards, prize), “agentic” (e.g., accomplish, leader, competent), “research” (e.g., scholarship, publications, grants), and “standout adjectives” (e.g., exceptional, outstanding, excellent). We developed two categories that reflect “positive evaluation” (e.g., groundbreaking, solid, comprehensive) and “negative evaluation” (e.g., illogical, unsubstantiated, diffuse) of a grant application” (Kaatz et al 2015). While this paper is less applicable to our research directly, the categories they establish will be extremely useful.

Data:

The data we will be drawing from comes directly from the University of Oregon's student evaluation system, as well as faculty data available via the registrar. The University of Oregon's registrar tracks all instructors' ranks. Additionally, near the end of each term, the registrar prompts students to complete course evaluations. The evaluation consists of twelve general questions and up to twenty-five questions departments can elect to add. Our analysis will be focused primarily on general questions eight and nine. Question eight reads "Please comment on the instructor's strengths and areas for possible improvement" and question nine reads "Please comment on the strengths and areas of possible improvement for the course as a whole". These are the only two general questions present across all courses where students are not constrained by a multiple-choice selection or by a numerical scale. The evaluation data includes the text data from both questions, CRN, term date, the course name and subject, and the instructor name. We will be evaluating data from Fall and Winter term (Terms 1 and 2) of the 2016-2017 academic year. We merge this data with our faculty data set containing professor characteristics such as gender and instructor rank to facilitate tests for bias and look for differences between different levels of teaching experience and tenure. After merging, we dropped any observation without a coded gender for the instructor.

In order to generate the emotionality score data that we needed for our regression analysis, we uploaded the text of the student evaluation surveys into LIWC. We put the evaluation data through LIWC, analyzing questions eight and nine separately per term. Using an internal dictionary, LIWC categorizes individual words and phrases contained within each text evaluation, then outputs the percentage of the total text that is found in said dictionary. Through this process LIWC provided us with numerical scores for the positivity, negativity and

emotionality of the evaluation texts. Therefore we can provide simple descriptive statistics based on the robust sentiment analysis systems built into the LIWC program. We then use these statistics to compare how instructors with different traits are written about by students. Our data set contains 54,075 observations (28,613 from fall term and 25,462 from winter term). Each observation is a student evaluation, unique per CRN, who answered question eight or nine.

We aggregated these student evaluation responses, attaching generated variables for the mean of positive and negative emotion, as well as our other descriptive outputs from LIWC to each independent CRN. This effectively gave us a single number for the mean positive and negative LIWC scores per class. Ultimately, this process allowed us to sort instructors by gender and rank, as well as differentiate between course subjects. 1,162 CRN's had no coded gender for the instructor and were subsequently dropped. We ran our regressions using the remaining 2,981 unique CRN's. Our final data set contains instructor name, instructor rank and gender, CRN, term, subject, college, average sum of word count for q8 and q9, means and standard deviations of LIWC output variables (positive emotion, negative emotion) for q8 and q9, and dummy variables for instructor rank and gender.

An example from our data of a student text response that LIWC would code with a high positive emotion score is "Interesting and useful". LIWC read this with a positive emotion score of 66.67 because two of three words, or 66.67%, are present in LIWC's positive dictionary. Below is a table (Table 1) that shows examples of other student responses along with the outputs LIWC provided itivity and negativity. While these particular responses were not randomly selected, we feel they adequately demonstrate LIWC's sentiment quantification process.

Table 1: Student Response Examples and Corresponding LIWC outputs

Student text response for question 8	LIWC positive emotion output	LIWC negative emotion output
POSITIVE DOMINATING		
"Great enthusiasm and makes the class enjoyable"	42.86	0
"She was amazing"	33.33	0
"Very engaging, made class really fun and exciting"	37.5	0
"Great course. Professor demonstrated a great knowlegde of material"	22.22	0
"This course was engaging, challenging, intriguing, and all around amazing experience. My only wish for improvent is that it be somehow longer."	6.67	0
NEUTRAL		
"he is a nice teacher, but has somewhat sad emotions"	10	10
NEGATIVE DOMINATING		
"Worst online class I have ever taken. I will never take one through the univeristy again"	0	6.25
"the worst teacher I have ever had in my school life"	0	9.09
"The organization of this course was painfully poor. This is the lowest quality instruction I have received in four years at this university"	0	13.04
"very difficult and struggled the entire time"	0	28.57

Methodology:

Our analysis has two parts: a descriptive search for bias and general tone in the surveys using WordStat results, and differences in student responses based on instructor gender using linguistic inquiry and word count, or LIWC results. By separating the project into these distinct parts, we have the ability to describe the evaluations generally, and then focus more specifically on each question using regressions.

Previous literature on score-based instructor assessment identified some potential problems with bias that female instructors could face. This potential bias prompted the University of Oregon to create an instructor assessment mechanism that put a greater focus on written feedback rather than simply asking students to put a score on their class experience. Our descriptive statistics seek to check if a similar pattern of bias is present in written reviews and see if instructors with specific rank and gender traits face bias in written reviews. In order to create a metric for how students feel about an instructor, we needed to have a mechanism that

mined the sentiment from a large number of written instructor evaluations to generate a metric that could be compared across groups.

To create this mechanism, we used the text analysis software WordStat. Modern text analysis techniques create the opportunity for our examination of written instructor assessments to provide comparable descriptive statistics on text traits that could previously be difficult to quantify. For the sake of our analysis, one such text trait that we wanted to generate is the general positivity and negativity of the written responses to our instructor assessment questions to see if instructors with different traits are described by students in different ways. To generate these statistics, we used the WordStat Sentiment Dictionary as a base list of negative and positive words and phrases, then used that dictionary to compare the frequency of positive and negative words and phrases between different groups of instructors. “Negative sentiment is measured by using the following two rules: Negative words not preceded by a negation (no, not never) within three words in the same sentence. Positive words not preceded by a negation within three words in the same sentence. Positive sentiment is measured in a similar way by looking for positive words not preceded by a negation as well as negative terms following a negation.” (Pelabeau, 2016)

We chose to use the WordStat Sentiment Dictionary because it acts as a comprehensive general use amalgamation of multiple top sentiment dictionaries and is not limited to a specific professional or academic field. Our data set contains text full of words and phrases specific to academic fields and courses, so we needed a sentiment dictionary that was as broad as possible. The WordStat Sentiment Dictionary was “designed by combining negative and positive words from the Harvard IV dictionary, the Regressive Imagery dictionary (Martindale, 2003) and the Linguistic and Word Count dictionary (Pennebaker, 2007). The WordStat dictionary building

utility program was then used to expand its word list by automatically identifying potential synonyms and related words as well as any inflected forms." (Pelabeau, 2016) This combination results in a list of "more than 9164 negative and 4847 positive word patterns". (Pelabeau, 2016) Using this comprehensive dictionary set, we then used WordStat to find and compare the frequency of words in the negative and positive categories.

We used regression analysis to test the relationship between the emotional tone of written reviews and instructor traits. This analysis is structurally similar to what Ancell and Wu used in their 2017 paper by using the class, instructor, and student characteristics and controls that they found to be relevant. However, we will be regressing against the variables for the positivity and negativity of the written evaluations that we got from LIWC. Using this measure of written student sentiment, we regress LIWC output measures against independent variables including: instructor rank, subject, and instructor gender. These results provide insight into how the characteristics of the instructors and different subject codes impact the LIWC output that instructors receive. The two regressions below are the general format we used for all four.

Table 2: Variable names and descriptions

Variable Name	Definition
q8meanposemo	Mean LIWC output per CRN of positive emotion for question 8
q8meannegemo	Mean LIWC output per CRN of negative emotion for question 8
q9meanposemo	Mean LIWC output per CRN of positive emotion for question 9
q9meannegemo	Mean LIWC output per CRN of negative emotion for question 9
instrcm	Dummy for instructor gender 1= male, 0= female
full	Dummy for full professor rank 1= professor, 0= not
assist	Dummy for assistant professor rank 1= assistant professor, 0= not
assos	Dummy for associate professor rank 1= associate professor, 0= not

inst	Dummy for instructor rank 1= instructor 0= not
countt	Number of text responses per CRN
q8sumwc	total number of words for all text responses per CRN for question 8
q9sumwc	total number of words for all text responses per CRN for question 9
senior	Dummy for all senior instructors (base, 1, and 2) combined 1= senior instructor, 0= not
EC	Dummy for Economics subject code 1= EC course, 0= not
MATH	Dummy for Math subject code 1= MATH course, 0= not
HIST	Dummy for History subject code 1= HIST course, 0= not
ARCH	Dummy for Architecture subject code 1= ARCH course, 0= not
ENG	Dummy for English subject code 1= ENG course, 0= not

$$q8meanposemo_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\, instrcm_i + \boldsymbol{\beta}_2\, full_i + \boldsymbol{\beta}_3\, assist_i + \boldsymbol{\beta}_4\, assos_i + \boldsymbol{\beta}_5\, inst_i + \boldsymbol{\beta}_6\, countt_i + \boldsymbol{\beta}_7\, q8sumwc_i + \boldsymbol{\beta}_8\, senior_i + \boldsymbol{\beta}_9\, EC_i + \boldsymbol{\beta}_{10}\, MATH_i + \boldsymbol{\beta}_{11}\, HIST_i + \boldsymbol{\beta}_{12}\, ARCH_i + \boldsymbol{\beta}_{13}\, ENG_i + \alpha_i + \mu_i$$

$$q8meannegemo_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\, instrcm_i + \boldsymbol{\beta}_2\, full_i + \boldsymbol{\beta}_3\, assist_i + \boldsymbol{\beta}_4\, ssos_i + \boldsymbol{\beta}_5\, inst_i + \boldsymbol{\beta}_6\, countt_i + \boldsymbol{\beta}_7\, q8sumwc_i + \boldsymbol{\beta}_8\, senior_i + \boldsymbol{\beta}_9\, EC_i + \boldsymbol{\beta}_{10}\, MATH_i + \boldsymbol{\beta}_{11}\, HIST_i + \boldsymbol{\beta}_{12}\, ARCH_i + \boldsymbol{\beta}_{13}\, ENG_i + \alpha + \mu_i$$

*The only change between these regressions, used for question 8, and the ones for question 9 is a q9 prefix before the means and word count sums.

Using LIWC to generate descriptive statistics on the language used by students in their written evaluations combined with WordStat frequency analysis provides the opportunity for simple but enlightening comparisons of how instructors with different traits are written about by

students. This method also provided useful numerical metrics for emotional tone and sentiment. These statistics provided can be applied to the models developed by Ancell and Wu to assess the relationships between written evaluations, SET score, and instructor quality in future projects.

Results:

These descriptive statistics provide a general overview of how negative and positive emotions compared across instructor groups when looking at our entire body of textual data for both questions combined. In order to narrow our focus and look at this data from a slightly different angle, as well as provide a broader exploration of our regression analysis, we chose to break the text up and focus on each question individually. This allows us to analyze the impact our various descriptive variables have on the negative and positive metrics generated by LIWC. We ran separate regressions for positive and negative emotion because LIWC codes them independently. For example, the presence of a low negativity rate in a response does not necessarily correlate with a high positivity rating. A single comment could have both positive and negative words. Thus, each metric needs to be examined separately rather than viewed as a single measure of sentiment.

Using our measure of general sentiment discussed in the methodology section, we chose to compare the percent negative and percent positive output of two separate sets of instructor traits: instructor gender and instructor rank. First, we examined the positive and negative output for all instructors, including both genders and all instructor ranks to establish a baseline we could compare to specific instructor categorizations. For the all-inclusive data set we found that 10.94% written text consisted of positive words and phrases and 2.61% was negative. (Figure 1)

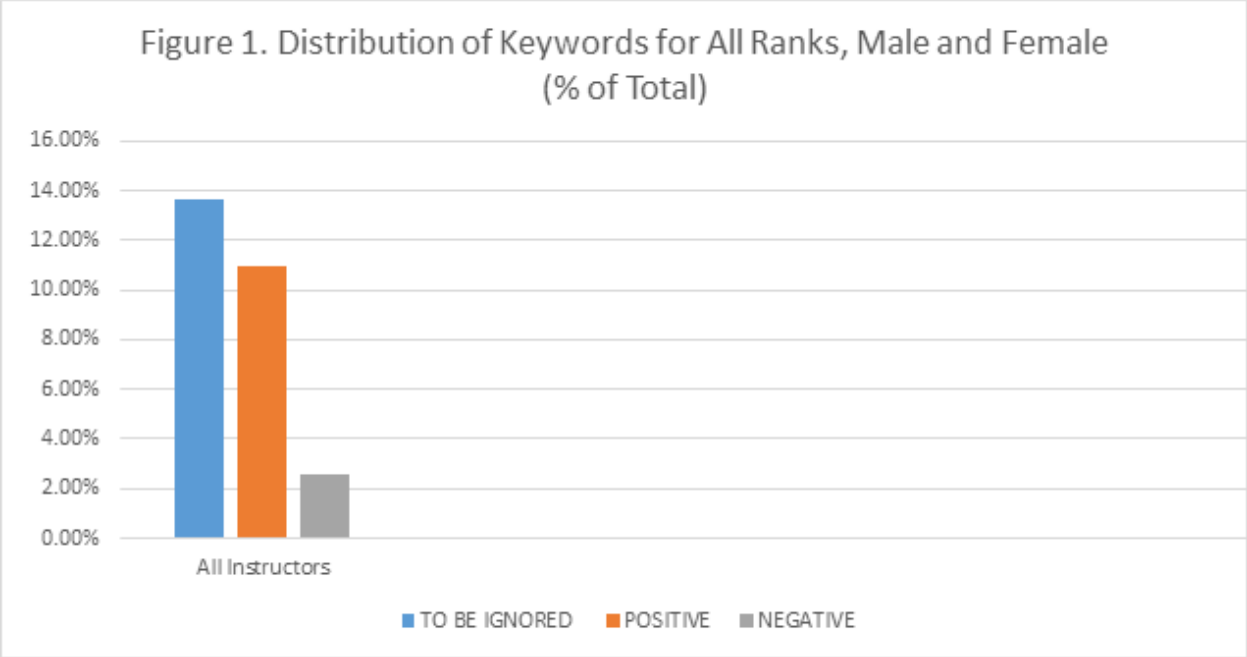


Figure 1. Male & Female, All Ranks (TO BE IGNORED: 314663 observations, 13.63% POSITIVE: 252477 observations, 10.94% NEGATIVE: 60316 observations, 2.61%)

For gender, we simply separated our data set into two separate sets, one for male and one for female, then generated positive and negative percent total for each category. For the all-female data set we found that 11.14% of the written text consisted of positive words and phrases and 2.52% was negative. (Figure 2) For the all-male data set we found that 10.76% of the written text consisted of positive words and phrases and 2.69%% was negative. (Figure 2)

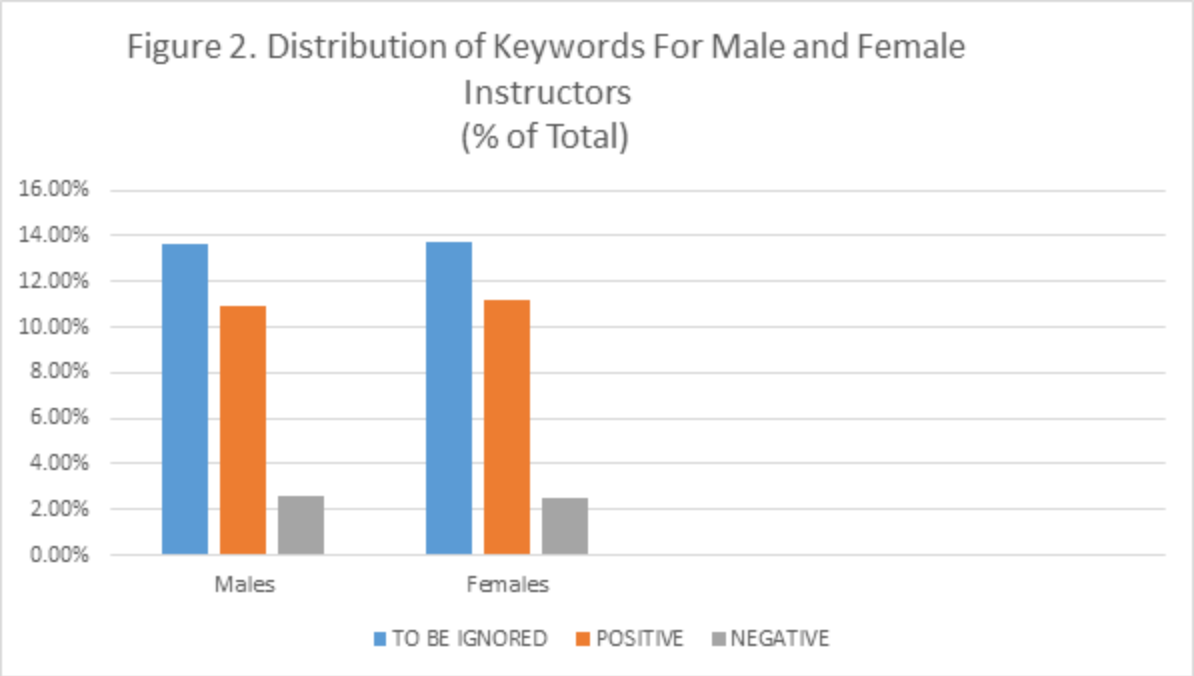


Figure 2. Male (TO BE IGNORED: 167750 observations, 13.53% POSITIVE: 133388 observations, 10.76% NEGATIVE 33337 observations, 2.69%) Female (TO BE IGNORED: 146913 observations, 13.75% POSITIVE: 119089 observation, 11.14% NEGATIVE: 26979 observations, 2.52%).

Based on this metric, it seems that female instructors have a slightly higher frequency of positive words and phrases in their evaluations when the two questions are examined at the same time. We used a similar method to examine the language used in question responses for instructors of different rank. We looked at three different instructor ranks: assistant professor, associate professor and full professor. These ranks can represent very different amounts of experience and often very different salaries, so it was a possibility that the instructor's rank impact student experience.

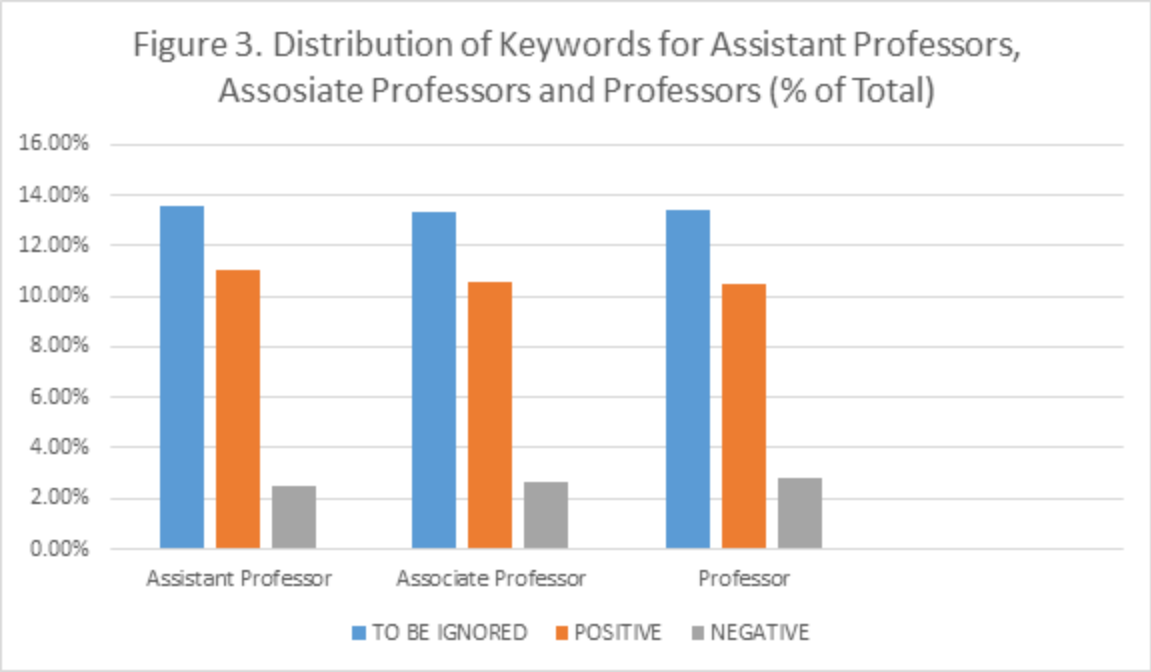


Figure 3. Assistant Professor (TO BE IGNORED: 45860 observations, 13.57% POSITIVE: 37240 observations, 11.02% NEGATIVE: 8371 observations, 2.48%). Associate Professor (TO BE IGNORED: 44776 observations, 13.32% POSITIVE: 35582 observations, 10.58% NEGATIVE: 8935 observations, 2.66%). Professor (TO BE IGNORED: 5866 observations, 13.44% POSITIVE: 5788 observations, 10.52% NEGATIVE: 3610 observations, 2.83%)

Based on this metric, it seems that students have generally similar sentiments about instructors from these different ranks. It is also worth noting that the percent positive and percent negative for each rank were similar to the percentages we found when looking at the genders individually. Going by the broad metric and looking at the text for both questions simultaneously, it does not appear that any single instructor trait is resulting in significantly more negative or more positive question responses.

WordStat also allows us to dig into the text and see what kinds of language was used to describe male and female instructors. This gave us the opportunity to generate descriptive

statistics on individual words used to describe instructors and courses and see if there are specific words or categories of words that are used more frequently to describe specific genders. To generate these statistics we took the 130 most frequently used positive and negative words from the evaluations and marked if a word was used more frequently to describe a particular gender at a statistically significant rate. (Figures 4 and 5) This list contains words that are used as few as 79 times in our entire data set. While such words may have little impact on the large trends of our data, the single use of a particularly strong negative or positive word can have a large impact on an instructor or an institution looking at written reviews of an instructor to review performance. This importance of word choice warranted the need to capture the gender trends of large numbers of positive and negative words.

Figure 4: Gender Designation of most frequently used positive and negative words (1st-66th most frequent)

Word	Frequency	Male, Female or Equal	Word	Frequency	Male, Female or Equal
Great	10519	m	Passion	850	equ
Good	9447	m	Approachable	814	f
Interesting	6147	m	Wonderful	807	f
Helpful	5953	equ	Awsome	806	equ
Feel	4080	f	Benefical	791	equ
Clear	3870	f	Effective	788	equ
Enjoyed	3716	f	Boring	779	m
Hard	3362	m	Exceptional	754	equ
Engaging	3285	m	Funny	732	m
Difficult	3032	m	Cares	729	f
Fun	2928	f	Interested	727	m
Easy	2744	m	Favorite	719	m
Passionate	2372	m	Problem	664	m
Knowledgeable	2153	equ	Enjoy	648	equ
Loved	2006	f	Valuable	640	f
Appreciated	1622	f	Comfortable	633	f
Amazing	1383	f	Recommend	615	equ
Engaged	1275	m	Excited	593	equ
Problems	1171	m	Positive	592	f
Love	1156	f	Enthusiasm	583	m
Excellent	1142	equ	Unclear	571	f
Relevant	1140	f	Professional	536	f
Loved	1158		Engage	527	f
Kind	1114	f	Fair	499	equ
Appreciate	1038	m	Confused	491	equ
Informative	1033	m	Interest	487	f
Enthusiastic	1018	equ	Order	488	equ
Confusing	1013	equ	Encouraging	484	f
Knowledgable	945	equ	Friendly	484	equ
Issues	919	f	Helping	475	m
Enjoyable	909	m	Opnions	470	f
Easier	901	m	Entertaining	466	m
Challenging	895	m	Perfect	446	equ

Figure 5: Gender Designation of most frequently used positive and negative words (67st-130th most frequent)

Word	Frequency	Male, Female or Equal	Word	Frequency	Male, Female or Equal
Frustrating	417	f	Stressful	200	equ
Incredible	406	f	Energetic	197	f
Harder	370	m	Meaningful	194	f
Fine	365	m	Efficiently	178	f
Waste	352	equ	Disappointed	177	m
Bad	351	m	Diverse	174	f
Unorganized	334	f	Fascinating	173	m
Exciting	332	equ	Poor	172	m
Care	329	m	Jokes	163	m
Patient	320	F	Complaint	159	m
Happy	320	m	Uncomfortable	158	f
Dry	318	m	Phenomenal	133	f
Cool	318	m	Poorly	133	m
Personable	309	equ	Impressed	129	f
Cared	290	f	Weakness	122	
Overwhelming	287	equ	Responsive	127	f
Relatable	287	f	Weakness	122	f
Vauge	284	m	Overwhelmed	122	f
Supportive	279	f	Weakness	122	f
Intellgent	274	equ	Negative	122	equ
Complaints	272	m	Dull	121	m
Tugh	269	m	Flexibility	119	f
Smart	265	m	Rude	111	m
Tough	269	m	Ridiculous	108	m
Smart	265	m	Humorous	94	m
Flexible	263	f	Comfort	94	f
Insightful	263	equ	Ideal	87	f
Looked	260	m	Nervous	86	f
Respectful	255	f	Joy	85	f
Inspiring	254	f	Condescending	84	f
Reasonable	247	equ	Sweet	83	f
Confident	229	f	Talented	79	equ

One notable finding about the rate specific words are used to describe different genders is that some of the largest gaps between males and females was between words that address an instructor's use of humor. The three most frequently used words that directly address an

instructor's use of humor are "funny", "jokes" and "humorous". These three words were all used around twice as often in instructor evaluations for males than females. The word "cool", which could likely be associated with instructors who use humor in their teaching, also had a large gender gap favoring male instructors. While female instructors did generally tend to have a slight advantage with extremely positive words like "inspiring", "sweet" and "joy", the gender gap was never nearly as large for these words. The largest consistent gender gap was for words having to do with humor.

Figure 6:"Funny" Rate per 10,000 Words (732 observations)

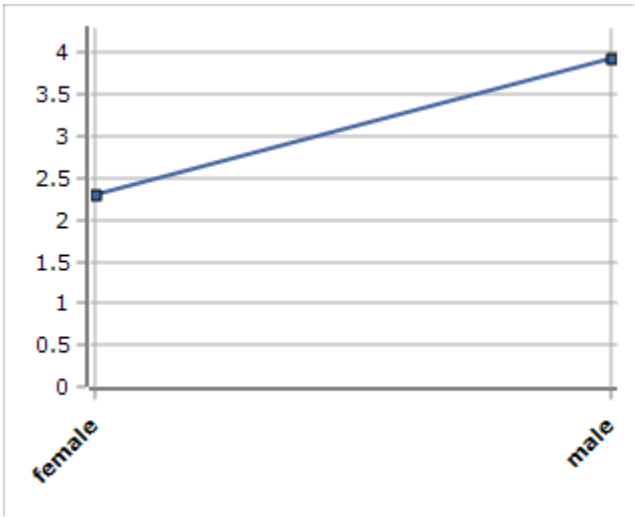


Figure 7: "Cool" Rate per 10,000 Words (318 observations)

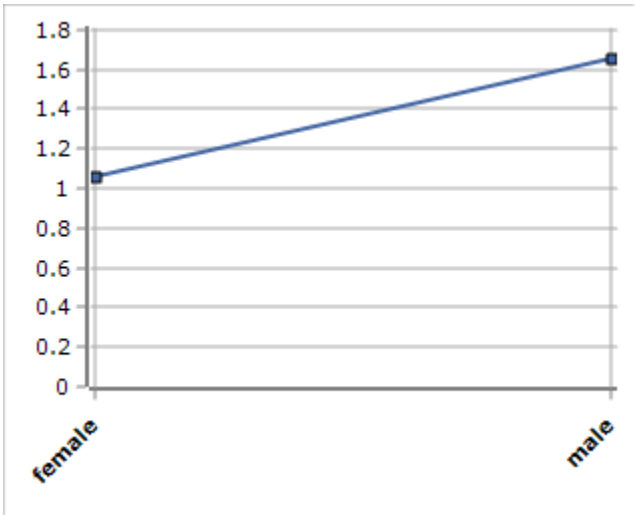


Figure 8: “Jokes” Rate per 10,000 Words (163 observations)

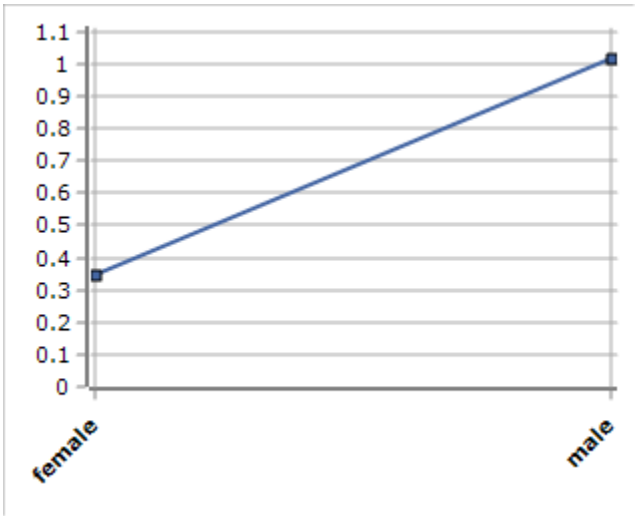
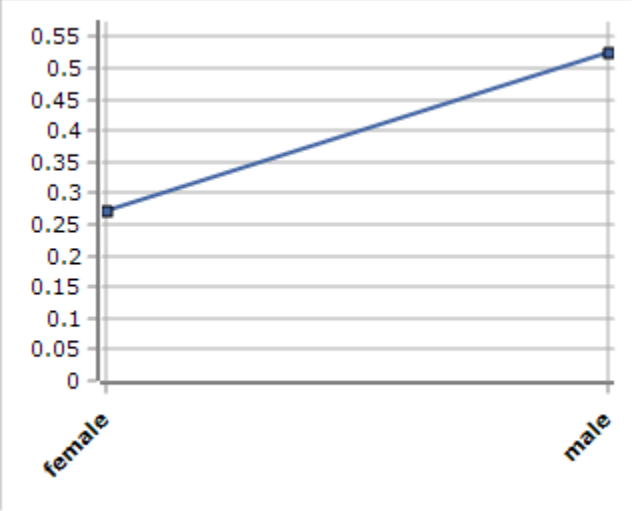


Figure 9: “Humorous” Rate per 10,000 Words (94 observations)



We also generated descriptive tables for the twenty most frequent words used to describe each gender. Our top 130 frequently used words list dives fairly deep into the full spectrum of words used by students. This simple comparison provides a more focused comparison of possible gender differences in the negative and positive words that are much more likely to appear in a given review. (Figures 10 & 11)

Figure 10: Twenty most Frequently Used Positive and Negative Words for Males

Word	Frequency	Word	Frequency
Great	5691	Easy	1516
Good	5273	Fun	1516
Interesting	3592	Pasionate	1300
Helpful	3208	Nice	1213
Clear	1945	Knowledgeable	1129
Hard	1887	Loved	906
Engaging	1859	Appreciated	776
Enjoyed	1850	Engaged	717
Organized	1752	Amazing	615
Difficult	1721	Excellent	599

Figure 11: Twenty most Frequently Used Positive and Negative Words for Females

Word	Frequency	Word	Frequency
Great	4828	Easy	1228
Good	4174	Loved	1100
Helpful	2745	Passionate	1072
Interesting	2555	Nice	1026
Clear	1925	Knowledgeable	1024
Enjoyed	1866	Appreciated	846
Organized	1721	Amazing	768
Hard	1475	Love	605
Engaging	1426	Kind	595
Fun	1412	Excellent	545

As shown above, the pool of most frequently used negative and positive words used to describe male and female instructors are very similar. This matches up well with our previous descriptive statistics and once again indicates that the language being used to describe each group is fairly similar based on the metrics we employed.

Figure 12: Positive Emotion Regression for Question 8

```
. regress q8meanposemo instrcm full assist assos inst countt q8sumwc senior EC MATH HIST ARCH ENG
```

Source	SS	df	MS	Number of obs	=	2,966
				F(13, 2952)	=	19.10
Model	6449.80238	13	496.138645	Prob > F	=	0.0000
Residual	76661.9374	2,952	25.969491	R-squared	=	0.0776
				Adj R-squared	=	0.0735
Total	83111.7398	2,965	28.0309409	Root MSE	=	5.096

q8meanposemo	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
instrcm	-.1494548	.1915312	-0.78	0.435	-.525003	.2260934
full	.0999085	.405441	0.25	0.805	-.6950672	.8948841
assist	.2019779	.4196573	0.48	0.630	-.6208727	1.024829
assos	.2899863	.3770943	0.77	0.442	-.4494082	1.029381
inst	1.660107	.3383904	4.91	0.000	.9966019	2.323612
countt	.1183476	.0119779	9.88	0.000	.0948617	.1418335
q8sumwc	-.0048666	.0004512	-10.79	0.000	-.0057513	-.0039819
senior	1.855173	.3755077	4.94	0.000	1.11889	2.591457
EC	-1.006474	.848544	-1.19	0.236	-2.670271	.6573244
MATH	-1.944851	.5019179	-3.87	0.000	-2.928996	-.9607069
HIST	.1317173	.9406917	0.14	0.889	-1.712761	1.976196
ARCH	-2.585239	.5108808	-5.06	0.000	-3.586957	-1.58352
ENG	.2846754	.4183633	0.68	0.496	-.5356379	1.104989
_cons	9.738502	.3060429	31.82	0.000	9.138423	10.33858

For our first regression we explored the relationship between the LIWC metric for the positiveness of the student response to question eight and a number of course and instructor traits (Figure 2). The total word count for all responses within a CRN (q8sumwc) has a statistically significant coefficient of -.00486, indicating that longer reviews are more likely to be negative. The department variable for courses in economics, architecture, and mathematics all had statistically significant and fairly large negative coefficients. The variable for math (MATH) having a coefficient of -1.944 and architecture (ARCH) having a coefficient of -2.585, strongly suggests that students are being the least positive about instructors in these two departments. For instructor rank, students tended to be the most positive about senior instructors and instructors relative to all other ranks, with the variable for senior instructor (senior) having a statistically significant coefficient of 1.855, and instructor having a statistically significant coefficient of 1.660. Our dummy variable gender (instrcm), where 1 indicates male and 0 indicates female has a statistically insignificant coefficient of -.1494.

Figure 13: Positive Emotion Regression for Question 9

```
. regress q9meanposemo instrcm full assist assos inst countt q9sumwc senior EC MATH HIST ARCH ENG
```

Source	SS	df	MS	Number of obs	=	2,934
				F(13, 2920)	=	14.77
Model	5031.60445	13	387.046496	Prob > F	=	0.0000
Residual	76504.9096	2,920	26.2003115	R-squared	=	0.0617
				Adj R-squared	=	0.0575
Total	81536.5141	2,933	27.7996979	Root MSE	=	5.1186

q9meanposemo	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
instrcm	.1007355	.1936649	0.52	0.603	-.2789981	.4804692
full	.7990494	.4105715	1.95	0.052	-.0059896	1.604088
assist	.3221435	.4219146	0.76	0.445	-.5051368	1.149424
assos	.8182026	.3808653	2.15	0.032	.0714108	1.564994
inst	1.539015	.3419733	4.50	0.000	.8684813	2.209548
countt	.0966048	.0104262	9.27	0.000	.0761613	.1170484
q9sumwc	-.0052826	.0004892	-10.80	0.000	-.0062419	-.0043234
senior	1.456029	.3788879	3.84	0.000	.7131145	2.198944
EC	.8650013	.8749461	0.99	0.323	-.8505726	2.580575
MATH	-.833035	.5023573	-1.66	0.097	-1.818046	.1519755
HIST	-1.169148	.9446311	-1.24	0.216	-3.021359	.6830624
ARCH	-1.823256	.5198581	-3.51	0.000	-2.842582	-.8039305
ENG	-.1417943	.4220422	-0.34	0.737	-.9693247	.6857362
_cons	7.050062	.3099507	22.75	0.000	6.442318	7.657807

For our second regression we used the same method on question nine to explore the relationship between the LIWC metric for the positiveness of the student response to question nine and the same set of course and instructor traits (Figure 3). Similar to question eight, responses by students taking a course in the architecture department were more likely to be negative, with the variable for architecture (ARCH) having a statistically significant coefficient of -2.585. Courses in the Math department fared slightly better than on question eight, with the variable for math (MATH) having a coefficient of -.8330 that it not quite statistically significant at the 95% confidence level. Senior instructor and instructor are again shown to be written about the most positively of all ranks, with the variable for senior instructors 1 and 2 combined (senior) having a statistically significant coefficient of 1.456, and instructor having a statistically significant coefficient of 1.539. Once again, total word count for all responses within a CRN (q8sumwc) has a statistically significant coefficient of -.00486, indicating that longer reviews are

more likely to be less positive. The coefficient on our dummy variable for gender is .1007 and is again statistically insignificant at the 95% confidence level.

Figure 14: Negative Emotion Regression for Question 8

```
. regress q8meannegemo instrcm full assist assos inst countt q8sumwc senior EC MATH HIST ARCH ENG
```

Source	SS	df	MS	Number of obs	=	2,966
				F(13, 2952)	=	11.58
Model	122.015884	13	9.3858372	Prob > F	=	0.0000
Residual	2391.93462	2,952	.810275956	R-squared	=	0.0485
				Adj R-squared	=	0.0443
Total	2513.95051	2,965	.847875381	Root MSE	=	.90015

q8meannegemo	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
instrcm	-.0279011	.0338317	-0.82	0.410	-.0942373	.0384351
full	.23717	.0716164	3.31	0.001	.0967469	.3775931
assist	.0844214	.0741275	1.14	0.255	-.0609254	.2297683
assos	.135681	.0666093	2.04	0.042	.0050756	.2662863
inst	.0978022	.0597727	1.64	0.102	-.0193982	.2150025
countt	-.0083467	.0021158	-3.95	0.000	-.0124952	-.0041982
q8sumwc	.0004902	.0000797	6.15	0.000	.0003339	.0006464
senior	.1643125	.066329	2.48	0.013	.0342566	.2943683
EC	.6893531	.1498853	4.60	0.000	.3954627	.9832434
MATH	.6141734	.0886579	6.93	0.000	.4403358	.7880109
HIST	-.0381174	.1661621	-0.23	0.819	-.3639228	.287688
ARCH	.0380045	.0902411	0.42	0.674	-.1389374	.2149463
ENG	.0015827	.073899	0.02	0.983	-.143316	.1464814
_cons	.4405815	.0540589	8.15	0.000	.3345846	.5465784

For our third regression, we looked at the relationship between the same set of instructor and class traits we used in the first two models and the negative sentiment metric generated by LIWC for the responses to question eight (Figure 4). Gender is again not statistically significant at the 95% confidence level and relatively small, with our gender dummy variable (instrcm) having a coefficient of -.0279. Longer responses are once again generally shown to be more negative, with the variable for entry word count (q8sumwc) having a statistically significant coefficient of .0004. The coefficient for our variable for full professors (full) is .2371 and statistically significant, meaning that of all the ranks full professors seem to get the highest concentration of negative words and phrases. Our variables for the math and economics departments both had statistically significant coefficients, with math (MATH) having a

coefficient of .6147 and economics (EC) .6893, making economics and math the two departments with the highest concentration of negative words and phrases for question eight.

Figure 15: Negative Emotion Regression for Question 9

```
. regress q9meannegemo instrcm full assist assos inst countt q9sumwc senior EC MATH HIST ARCH ENG
```

Source	SS	df	MS	Number of obs	=	2,934
				F(13, 2920)	=	6.29
Model	120.968228	13	9.30524833	Prob > F	=	0.0000
Residual	4318.92772	2,920	1.47908484	R-squared	=	0.0272
				Adj R-squared	=	0.0229
Total	4439.89595	2,933	1.51377291	Root MSE	=	1.2162

q9meannegemo	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
instrcm	.113344	.0460145	2.46	0.014	.0231199	.203568
full	.0408285	.0975511	0.42	0.676	-.1504474	.2321044
assist	.0492212	.1002462	0.49	0.623	-.1473392	.2457816
assos	.3108602	.0904929	3.44	0.001	.1334237	.4882966
inst	.120631	.0812523	1.48	0.138	-.0386866	.2799485
countt	-.002925	.0024773	-1.18	0.238	-.0077823	.0019324
q9sumwc	.0003433	.0001162	2.95	0.003	.0001154	.0005712
senior	.1643738	.0900231	1.83	0.068	-.0121415	.340889
EC	.3972223	.2078857	1.91	0.056	-.0103951	.8048398
MATH	.6845777	.1193592	5.74	0.000	.4505409	.9186145
HIST	.1186362	.2244427	0.53	0.597	-.3214459	.5587183
ARCH	-.0110406	.1235174	-0.09	0.929	-.2532306	.2311495
ENG	.1029708	.1002765	1.03	0.305	-.0936491	.2995906
_cons	.4959312	.0736438	6.73	0.000	.3515322	.6403302

Our fourth and final regression looks at the same set of instructor traits in relation to the LIWC metric for negative sentiment in the text responses for question 9 (Figure 5). Word count remains a consistent indicator of sentiment, with our word count variable (q9sumwc) having a statistically significant coefficient of .0003. The only statistically significant rank trait was associate professor, with our variable (assos) having a coefficient of .0311. Math was the only department that was statistically significant, with our variable (MATH) having a coefficient of .6845.

Discussion:

The descriptive portion of our analysis consistently demonstrated that the general rate of negative and positive word and phrase use is similar between all groups of instructors we examined. Our method for measuring, quantifying, and comparing the sentiment of the text written about our various instructor groups was fairly straight forward, allowing us to apply our method consistently across all groups and generate easily comparable results. One of the most pressing concerns for institutions trying to improve instructor evaluation methods is the potential for gender bias. With our sentiment metric finding only a 0.22 percent gender difference in positiveness and a 0.17 percent gender difference in negativity, we did not find any evidence that could confirm a hypothesis that written instructor evaluations favor one gender over another with this particular metric. Based on the relatively small but potentially problematic degree of bias described by Ancell and Wu in their analysis of numeric instructor evaluations, this suggests that written reviews at the University of Oregon potentially present fewer problems with gender bias than the numeric scores.

It is also interesting to note that while these differences were too small to make any definitive claims about gender bias, female instructors did maintain a slight edge in positivity over male professors. This is different that what Ancel and Wu found in their exploration of numeric instructor evaluations, where all possible bias was to the advantage of male instructors and to the disadvantage of female instructors. While our descriptive statistic model did not find extensive evidence of gender bias that lead to one set of evaluations being more positive or negative than another, we did find some trends that suggest the language used to describe instructors may change on the basis of gender. While the set of the top twenty positive and negative words used in instructor evaluations were remarkably similar, words describing humor

were used around twice as often in evaluations of male instructors. None of the words having to do with humor were used at a very high rate, so it would be inaccurate to suggest that this phenomenon was necessarily impacting the overall performance measure of a large portion of instructors. Still, this trend does suggest that a metric less focused on positive and negative sentiment and more focused on the different ways male and female instructors are written about by students could yield further insight into the potential trends of bias within a written evaluation system. It is difficult to say how text traits like the mention of a good sense of humor could impact a given instructor's chances of receiving a pay raise or promotion, but if other, similar trends are found in the future it could illuminate other advantages one group of instructors may have over another.

Our descriptive statistics also found little difference in overall negative and positive sentiment between the different instructor ranks. The largest contrast in our measures of positive and negative sentiment was between assistant professors and full professors, with assistant professors having 0.5 percent more positive and 0.35 percent fewer negative words and phrases in their evaluations than full professors. While this is the largest gap in sentiment we found between two comparable groups of instructors with our descriptive model, it remains small enough that this metric alone is not enough to suggest that one group of instructors has a significant advantage over another on its own. This descriptive result is consistent with Ancell and Wu's finding that full professors were rated statistically significantly lower with the numeric scores. Our metrics for negative and positive sentiment are not entirely comparable with the numeric scores examined by Ancell and Wu, but our results show that the trend of full professors receiving more negativity and less positivity is present for both written and numeric instructor evaluations to some extent.

The results from our regressions fail to prove the existence of statistically significant gender difference. There was no clear indication that either gender had significantly different LIWC scores for positivity or negativity. For example, our regression on the mean positivity of question eight yielded a negative coefficient on our gender dummy variable, suggesting male instructors were written about less positively. This result was similar across the four regressions, each coefficient on gender was small and statistically insignificant, therefore we cannot establish with confidence that the text contains bias. The reverse of this finding is that the text from student evaluations is relatively better than numerical scores as a source for conducting faculty reviews. Additional analysis into this area is necessary to answer this question with more conclusivity.

The most robust results in this regression (Figure 12) are the coefficients on rank. Statistically significant and large positive coefficients on all levels of instructor (instructor, senior instructor, senior instructor 1, and senior instructor 2) suggest that non-tenure track teaching staff receive more positive student evaluations. This could be due to instructors and senior instructors having fewer department and research responsibilities, leading to more time for student engagement and communication. Our regression on the mean negativity of question eight found that full professors receive higher negativity scores than their peers. This result reinforces the conclusions made about non-tenure track faculty made by Ancell and Wu.

Across all of our regressions, the variable for number of student evaluations in a class (countt) demonstrated that more evaluations correlated with higher positivity and lower negativity scores. This finding was statistically significant in three out of four regressions. Interestingly however, a greater total amount of words that students wrote in an evaluation answer (q8sumwc and q9sumwc) correlated with lower positivity and higher negativity scores.

This is consistent and statistically significant in all four regression models. The inverse relationship between word total and evaluation count suggests students are writing a greater amount of shorter evaluations for courses they enjoyed or thought were conducted positively, and fewer, longer evaluations for courses they had negative experiences in. This result aligns with anecdotal stories of staff members receiving isolated, distinctly negative targeted student evaluations that in part lead to the creation of this project.

Another useful variable analyzed in our models was course subject. We found that math, economics, and architecture all had statistically significant coefficients present across several regressions. Math consistently had a large correlation with decreased positivity and increased negativity, statistically significant in three of the four regressions. Likewise, our variable representing architecture had highly negative coefficients, statistically significant in both regressions on mean positivity. Economics had a statistically significant coefficient for the question eight regression on mean negativity. The coefficients indicate that these subjects receive consistently less positive and more negative evaluations. For math and architecture, the evidence is strong for responses about both course and instructor feedback. Math, economics, and architecture are all quantitative subjects and almost universally described by students and staff as the most difficult courses at the University of Oregon. The direction of these coefficients supports the hypothesis Ancell and Wu put forward about the positive relationship between grades and course feedback. These courses tend to have lower average class GPA's, so it follows that students' overall decreased satisfaction with their class would be reflected in the text evaluation as well. Future projects focusing on the contrasting between these courses, and qualitative subjects like humanities and art, could better illuminate the differences in text evaluations for their respective faculty.

Conclusion:

Taken as a whole, the two models presented here for analyzing written student evaluations identify compelling trends present across our data set. The positive and negative sentiment of the evaluations were generally similar for both genders. When analyzing our entire data set of written evaluations we did not find compelling evidence that a given individual staff member is at a serious disadvantage due to gender bias. We did find that full tenure track professors received measurably more negative and less positive evaluations, but the actual impact rank made on the overall sentiment was relatively small. This does not mean that individual instructors cannot face bias or discrimination in their written reivews because of these traits. Our paper focused on the measured differences between instructor and course traits; nevertheless, it is worth noting that the vast majority of total language across all evaluations was positive, indicating an overall positivity bias.

We conclude that based on our metrics there were no identifiable trends suggesting that the use of written evaluations presents widespread structural problems with bias that would make them unfair for an institution to use when evaluating instructor performance. At the very least, written reviews offer more opportunity for case by case interpretation for institutions examining instructor performance, and perhaps leave more room for students to better describe their experience with both the course and the instructor. A written evaluation has the potential to provide significantly better context for the feelings a student has about a given course. Written evaluations can also make any bias that a student has when evaluating an instructor more apparent. This allows institutions seeking an accurate assessment of instructor performance a greater ability to adjust their interpretation of student feedback if significant degrees of bias are

found to be in the text. We see this paper as a starting point for text responses to be further explored as a better alternative to numerical systems for instructor evaluation.

Works Cited

Ancell , Kenneth, and Emily Wu. "Teaching, Learning, and Achievement: Are Course Evaluations Valid Measures of Instructional Quality at the University of Oregon? ." *University of Oregon*, June 2017.

Fan, Mingming and Maryam Khademi. "Predicting a Business Star in Yelp from Its Reviews Text Alone." *CoRR* abs/1401.0864 (2014): n. pag.

Kaatz, Anna, et al. "A Quantitative Linguistic Analysis of National Institutes of Health R01 Application Critiques From Investigators at One Institution." *Academic Medicine*, vol. 90, no. 1,Jan. 2015, pp. 69–75., doi:10.1097/acm.0000000000000442.

Kaatz, Anna, et al. "Analysis of National Institutes of Health R01 Application Critiques, Impact, and Criteria Scores." *Academic Medicine*, vol. 91, no. 8, Aug. 2016, pp. 1080–1088., doi:10.1097/acm.0000000000001272.

Pak, Alexander & Paroubek, Patrick. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of LREC. 10.

"Revising UO's Teaching Evaluations." *Office of the Provost*, University of Oregon, Oct. 2018, provost.uoregon.edu/revising-uos-teaching-evaluations

Pelabeau, Normand. "Sentiment Analysis with WordStat." *Provalisresearch.com*, Provalis Research , 2016, provalisresearch.com/uploads/WP_SentimentAnalysis_LR.pdf.