Stanford cepa | Center for Education Policy Analysis

CEPA Working Paper No. 19-04

Text as Data Methods for Education Research

AUTHORS

Lily Fesler
Stanford University

Thomas Dee
Stanford University

Rachel Baker
University of California, Irvine

Brent Evans
Vanderbilt University

ABSTRACT

Recent advances in computational linguistics and the social sciences have created new opportunities for the education research community to analyze relevant large-scale text data. However, the take-up of these advances in education research is still nascent. In this paper, we review the recent automated text methods relevant to educational processes and determinants. We discuss both lexical-based and supervised methods, which expand the scale of text that researchers can analyze, as well as unsupervised methods, which allow researchers to discover new themes in their data. To illustrate these methods, we analyze the text interactions from a field experiment in the discussion forums of online classes. Our application shows that respondents provide less assistance and discuss slightly different topics with the randomized female posters, but respond with similar levels of positive and negative sentiment. These results demonstrate that combining qualitative coding with machine learning techniques can provide for a rich understanding of text-based interactions.

VERSION

June 2019

Suggested citation: Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as Data Methods for Education Research (CEPA Working Paper No.19-04). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp19-04

I. Introduction

The education-research community is in the middle of two advances that are changing the way we can analyze and understand educational processes. First, more than ever in the past, we have access to broad, rich, educationally-relevant text data from sources such as online discussion forums, transcribed video and audio of face-to-face classes, digital essays, social media, and emails and texts between parents and schools. Second, computational linguists and social scientists have recently developed more advanced and nuanced tools that can be applied to quantitatively analyze large-scale text data (Gentzkow, Kelly, & Taddy, 2017; Grimmer & Stewart, 2013; Jurafsky & Martin, 2018). Together, these advances in data availability and analytic techniques can dramatically expand our capacity for discovering new patterns and testing new theories in education. They can expand the types of research questions that researchers can ask, improve the external validity and representativeness of research, and reduce the cost to complete these types of studies.

Researchers can now ask questions that rely on previously untapped data sources, utilize more real-time data, and focus on mechanisms in addition to impacts. For example, researchers can utilize the discussion board forums of online classes to study classroom interactions, teachers' application essays to understand teacher motivation (Penner, Rochmes, Liu, Solanki, & Loeb, 2019), district web sites to catalog school improvement plans, or newspapers to study how educational policies are discussed in the media (Mockler, 2018). Instead of asking teachers to recall any professional development opportunities they have utilized, researchers can directly study the topics that teachers discuss in online professional development communities in real time (Anderson, 2018). And, in addition to studying the impacts of a text messaging campaign, researchers can use the content of the text messages to understand how and why a program worked in the way that it did (Fesler, 2019).

Researchers can also use new data mining methods to discover new codes or themes in their data when conducting content analyses. For example, Reich, Tingley, Leder-Luis, Roberts and Stewart (2015) used automated text analysis to discover students' motivations in online classes without imposing their

own groupings or pre-determined themes. These methods can be useful in generating new theories based on the discovery of new topics (Nelson, 2016).

In this paper, we review some of the recent automated text methods that researchers can apply to educational processes and determinants. We discuss both lexical-based and supervised methods, which expand the scale of text that researchers can analyze, as well as unsupervised methods, which allow researchers to discover new themes or groupings in their data. We also provide recent examples in the education literature that employ these methods. To illustrate these methods further, we apply them to data from a randomized experiment in which fictitious students with gendered names post comments in the discussion forum of 124 Massive Open Online Courses (MOOCs). Analyzing student and instructor responses to students' comments using the methods we discuss allows us to examine the causal impact of a student's perceived gender on the language of instructor and student responses.

We use lexical-based methods to examine whether posts differ in positive and negative sentiment, supervised machine learning to examine whether posts differ in confirmation (i.e. assistance, acknowledgement individual attention, and disconfirmation), and unsupervised machine learning to examine whether posts differ in topics discussed. We show that respondents provide 7 percentage points less assistance and 7 percentage points more individual attention to (the black and white) female posters. Students' and instructors' provision of less informational feedback to female students is likely to be educationally salient in that it may discourage dimensions of engagement among female students. Through an exploratory analysis, we also observe that respondents discuss slightly different topics with male and female students. Through this case study, we illustrate the great promise, and potential pitfalls, of textual analysis methods.

II. An Overview of Automated Text Analysis Methods

The existing methods for analyzing text generally cluster into three categories: lexical-based methods, supervised machine learning, and unsupervised machine learning. Lexical-based methods rely on simple document-level[1] word counts to accomplish tasks as diverse as measuring sentiment in a student essay, determining whether an online discussion post was directed at a particular student (Bettinger, Liu, & Loeb, 2016), and measuring policy uncertainty in newspaper articles (S. R. Baker, Bloom, & Davis, 2016). Supervised machine learning relies on researchers' hand-coded documents to train a model and predict the codes for an unlimited number of un-coded documents, which can dramatically expand the number of documents that a researcher can code. For example, Kelly, Olney, Donnelly, Nystrand and D'Mello (2018) use supervised machine learning to predict when teachers ask questions without predetermined answers by building a model based on a set of hand-coded questions. Similarly, Gegenheimer, Hunter, Koedel and Springer (2018) hand-code teacher observation feedback into seven major domains for a subset of teachers, creating a foundation for using supervised machine learning to predict the domains for all teachers in Tennessee. Unsupervised machine learning procedures instead identify groupings in the data without relying on prior hand-coding, which is a less structured approach that can be particularly appropriate for researchers in a more descriptive and hypothesis-generating posture. For example, Beattie, Laliberte and Oreopoulos (2018) use unsupervised machine learning to examine open survey responses of students who perform significantly better or worse than predicted in their first year of college. They find that students who perform lower than expected are more likely to discuss topics like "getting rich" quickly, and students who perform better are more likely to discuss more philanthropic goals.

[1] In keeping with the literature, we use the term "document" to refer to one observation of text (e.g. one essay, one text message, or one discussion board post).

When deciding which of these methods is appropriate for a particular analysis, it is important to have a crisp understanding of the motivating research questions. We use Figure 1 to outline when each of these three broad methods are most appropriate. If researchers intend to test a particular theory, they should use lexical-based or supervised machine learning methods. Lexical-based methods can be applied when researchers know the types of terms that make up their construct of interest (e.g. positive words would be an appropriate way to measure positive sentiment in text). Supervised machine learning methods can be applied when researchers do not know the specific terms that make up their construct but can easily code their documents by reading a subset of them. In contrast, researchers should use unsupervised machine learning if they are interested in discovering new themes in their data rather than testing a particular theory (i.e. exploratory rather than confirmatory inquiries). Topic modeling is a particularly useful way to measure new themes in data. We discuss each of three approaches in more detail below.

Insert Figure 1 here.

A. Lexical-Based Methods

If researchers know the types of terms that make up their construct, lexical-based methods work best. These methods rely on the number of times words occur in documents to measure the construct of interest. For example, Baker, Bloom, and Davis (2016) determine whether newspaper articles discuss policy uncertainty based on whether the articles contain terms like "regulation," "deficit," and "federal reserve;" Evans, Marsicano, and Lennartz (2019) assess postsecondary institutions' commitment to civic engagement by observing whether terms such as "volunteer" and "service" exist in the institutions' mission statements; and Bettinger, Liu, and Loeb (2016) measure peer interactivity in online classes by constructing a course-specific roster and determining whether each post contained a name on that course's roster or not. In each of these cases, the researchers were able to determine their lists of words from their own data.

However, despite knowing the types of terms that make up their construct, researchers may not always be able to easily construct a word list from their own data. In this case, researchers can consider using an off-the-shelf tool, known as dictionaries or lexicons. There are many of these tools available, including Linguistic Inquiry and Word Count (LIWC) and Sentiment Analysis and Social Cognition Engine (SEANCE). LIWC is the most widely used dictionary-based sentiment analysis tool used among social scientists, and its word lists were constructed through a combination of using existing dictionaries, thesauruses, common emotion rating scales, and human-generated lists, and ratings by judges (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Each of the words on the LIWC lists are weighted equally. SEANCE is a newer tool that combines lists of words from eight open source databases with tools that identify parts of speech and negations. This allows SEANCE to analyze adjectives separately from nouns (for example), and to ignore terms with negations (e.g. not) before them (Crossley, Kyle, & McNamara, 2017). SEANCE also includes summary indices produced from a principal component analysis on a corpus of movie reviews from the Internet Movie Database (IMDb) (Crossley et al., 2017; Pang & Lee, 2004). Each of the words from SEANCE are thus weighted based on their principal component loadings.

The most common types of dictionaries measure positive and negative sentiment, but there are also many other types of dictionaries that measure whether documents contain content as diverse as politics, biological processes, time orientations, and evaluative adjectives (Pennebaker et al., 2015; Stone, Dunphy, Smith, & Ogilvie, 1966). Lexical-based tools like SourceRater and Coh-Metrix can also measure constructs like text complexity (Burstein, Sabatini, & Shore, 2014; Graesser, Mcnamara, & Kulikowich, 2011).

Despite the ease of use of external dictionaries (and their correspondingly wide application), they can perform poorly when applied to a different domain. For example, Loughran and McDonald (2011) showed that almost three-fourths of the negative terms in generic dictionaries (like tax, cost, liability, and vice) have positive or neutral connotations in earnings reports (which is their subject of interest). Goncalves et al. (2013) also showed that sentiment analysis tools can vary substantially in their ratings of

the positivity of texts. Thus, researchers should recognize the risks associated with external dictionaries and seek to validate their use for a specific sample and context.

The best way to validate dictionaries is to compare the dictionary methods to hand coding.[2] For example, if using a sentiment dictionary that categorizes documents as positive or negative, researchers can hand code a subset of their documents as positive or negative. Researchers can then compare the dictionary measures to their hand coding using a table known as a *confusion matrix*, which contains four boxes: true positives (the computer correctly categorized a document as positive), false negatives (the computer incorrectly categorized a document as negative), false positives (the computer incorrectly categorized a document as positive), and true negatives (the computer correctly categorized a document as negative).

This matrix can be summarized using three measures: *recall*, *precision*, and the *F1-measure*. Recall measures the percentage of the hand coded positive instances that were predicted by the computer as being positive, precision measures the percentage of the computer's positive predictions that were hand-coded as being positive, and the F1-measure is the harmonic mean of precision and recall.[3] Researchers would ideally like to maximize both recall and precision, but they can also choose to minimize the type of error that would be most problematic for their particular application. It should also be noted that researchers can also validate their measure by comparing their dictionary measure to other measures that represent similar constructs. For example, Quinn et al. (2010) correlate their measure of Congress' discussion of abortion to official abortion roll-call votes, and Baker et al. (2016) correlate their measure of economic policy uncertainty to an established measure of volatility in the S&P500 index.

---

[2] Hand coding is a traditional method frequently employed in qualitative research and content analysis in which a person reads and manually codes words or phrases with specific themes.

[3] The harmonic mean is the reciprocal of the arithmetic mean of reciprocals and is more conservative than using the arithmetic mean (i.e. it produces a lower F1-measure).

B. *Supervised Machine Learning*

If researchers know what theory they would like to test but do not know which words or linguistic features make up their construct, supervised machine learning is a better option. This works well when humans can relatively easily categorize their data by reading each document, but there are too many documents to read in a reasonable amount of time. These methods involve hand-coding a subset of the total documents into the categories of interest, then using the relationships between document-level features and their hand-coded categories to predict the categories for unlabeled documents. For example, Sajjadiani, Sojourner, Kammeyer-Mueller and Mykerezi (2019) use supervised machine learning to classify teacher applicants' self-reported reasons for leaving their previous job into four categories (involuntary, avoiding bad jobs, approaching better jobs, and other reasons). They hand-coded 1,000 of the self-reported reasons, then were able to train a model using those 1,000 observations to predict the reasons for the remaining 35,000 applicants. They could then use these predictions to study the relationship between teacher turnover reasons and job performance.

In supervised machine learning, the researcher begins by hand-coding a subset of documents for their construct(s) of interest. This is often known as *labeling* the data. For example, researchers may want to predict whether a particular post is positive or negative. Just as in qualitative coding, the reliability of the hand-coded labels should be measured, typically by having multiple people code the documents and estimating inter-rater reliabilities. Researchers should construct a set of *features*, or variables, that they think may be predictive of their hand codes. In many instances, researchers will simply include indicators for whether each word occurs in a given document, and this set of variables is known as a *document-term matrix* (since documents serve as rows and terms as the columns; see Section IID). However, they may also choose to include additional features, like parts of speech, the length of the document, or even lexical-based measures to improve predictive power.

Researchers then build their model using one subset of the data (the *training* data) and assess model performance on another subset of the data (the *test* data). This is often done using k-fold cross

validation (Grimmer & Stewart, 2013). Researchers randomly divide their hand-coded data into k subsets (folds) of equal size, then build their model using k-1 folds and assess performance on the held-out fold. They repeat this process k times, until they have a prediction for each hand-coded observation. Researchers can then compare the predictions for each observation to the hand-coded data to assess model performance. They would then apply the model that minimizes the test error to the data that has not been hand-coded (Hastie, Tibshirani, & Friedman, 2009, Chapter 7).

Researchers should consider using multiple types of machine learning algorithms to assess which maximizes performance. Researchers could consider ordinary least squares or logistic regression; regularized regression like LASSO, ridge, or elastic net (Hastie et al., 2009, Chapter 3); Support Vector Machine (SVM) (Hastie et al., 2009, Chapter 12); decision tree algorithms like random forests (Breiman, 2001; Hastie et al., 2009, Chapter 15); or a deep learning method like a neural net (Hastie et al., 2009, Chapter 11). Taking the example of logistic regression, the researcher would regress their positive hand-coded measure on their features (e.g. their document-term matrix), which would produce a set of positive and negative weights associated with each of their features (e.g. terms). For example, there might be a positive weight on the term "great" and a negative weight on the term "terrible." They could then use these weights to predict whether a particular observation in their test data (or in their held-out fold in k-fold cross validation) is positive or negative and assess how the prediction compares to the true value. The other supervised machine learning methods similarly build a predictive model based on the relationship between their hand-coded measure and their features.

In addition to producing document-level classifications, supervised machine learning can also be used to estimate overall proportions. For example, suppose Sajjadiana et al. (2019) were interested in the proportion of teachers who leave voluntarily from their previous job. An issue to confront is the need to adjust for any bias introduced into the overall proportions from small biases in the individual-level predictions. The Hopkins-King method can be used to correct for this. For example, the model that minimizes individual document misclassification may mis-categorize 10% of the leave codes as "avoiding bad jobs" instead of "involuntary leaving." This model, which is 90% accurate, may perform well overall.

However, the overall proportion for "avoiding bad jobs" is 10 percentage points too high, and for "involuntary leaving" is 10 percentage points too low. Researchers cannot correct for this on an individual document basis, but they can adjust the overall proportions to be 10 percentage points higher or lower as needed (Hopkins & King, 2010; Jerzak, King, & Strezhnev, 2018).[4]

C. Unsupervised Machine Learning

If researchers would like to discover new theories or topics in their data, unsupervised methods are preferred. Unsupervised methods automatically find topics or clusters within the data, and group documents (or parts of documents) into those groups by modeling a text's word choice. These methods can vary by whether documents can contain single or multiple topics (i.e. single membership versus mixed membership models), and whether the topics are fully automated or computer assisted[5] (Grimmer & Stewart, 2013). One of the unsupervised methods that would be of most use to social science researchers is topic modeling, which is a fully automated mixed membership model (Blei, Ng, & Jordan, 2003). Penner, Rochmes, Liu, Solanki, and Loeb (2019) use topic models on essay responses of 10,000 teachers to examine how teachers perceive their role in closing achievement gaps. Topic modeling allowed them to examine essay content without imposing their own groupings on the data and led to the discovery that that Hispanic and African American applicants discuss structural causes of inequality more frequently. Reich, Stewart, Mavon, and Tingley (2016) use topic models to explore liberal and conservative students' language in MOOCs, and find that these groups of students use largely the same language to discuss similar topics (like the Common Core and school vouchers).

---

[4] The R package "readme2" implements this; see https://github.com/iqss-research/readme-software.

[5] With computer assisted clusterings, the researcher can explore many possible computer-generated clusterings (Grimmer & King, 2011).

Intuitively, topic models use the terms in the documents (which are observable) to determine the topics being discussed (which are latent and unobservable). The data generating process can be conceptualized in the following way. First, the author of each document decides which topics they would like to discuss. Second, they select the words to include in their response, and are more likely to select words that are more highly associated with that topic. They continue selecting words to create their bag-of-words response. The objective of the topic model is to find the parameters that have generated these bag-of-words documents.[6]

Topic modeling can also be used to explore how content varies by document characteristics, like characteristics of the document's author. This model is known as a structural topic model (STM), and it allows topic prevalence to vary by specified covariates. It estimates the topics and the relationships between the topics and the covariates simultaneously, which allows the model to propagate the uncertainty from the topic proportions into the estimation of the regression coefficients (Roberts et al., 2014). As an example, Reich et al. (2015) used a structural topic model to estimate how student motivations in online courses vary by gender, and found that male students were more likely to state that they enrolled in a class because of the university's elite association than female students were.

After estimating the topic model, the researcher examines the topic groupings and manually labels the topics. To do so, they first examine the words that are most frequent and exclusive for each topic (e.g. using FREX)[7] and determine labels for each topic. For example, Reich et al. (2015) find that one of their topics contains terms like "improve," "skill," "develop," "career," and "profession," and label that topic "Professional Development." To validate these labels, the researcher should then read a subset

---

[6] We should note that topic models are multi-modal, meaning that topics can be sensitive to the starting values used. One way to account for this is using a spectral initialization, which is deterministic and globally consistent (Roberts, Stewart, & Tingley, 2016).

[7] The FREX measure was developed by Roberts, Stewart, and Airoldi (2016), and is the weighted harmonic mean of the word's rank in terms of exclusivity and frequency.

(often 10-20) of the documents that are mostly highly associated with each topic and determine 1) whether their labels need to be refined, and 2) whether the topics cohere well (Quinn et al., 2010).

D. Pre-processing text

All approaches begin with transforming the text under consideration into quantifiable measures. Specifically, researchers begin with a *corpus* (i.e., collection of texts), which consists of their documents. For example, in the case study we present below, each response to a post in the discussion forum of a MOOC course is a document, and all of the responses collectively make up our corpus.

The methods discussed in this article commonly treat each document as a "bag-of-words." That is, word order, punctuation, and capitalization are not taken into account. [8] The primary input to these types of analyses is the number of times a given word occurs in a document. To prepare text for this type of analysis, researchers typically convert their text to lowercase, remove all punctuation and numbers, and fix common misspellings.[9] Researchers also remove *stopwords* or functional words that occur very frequently but do not contain much content (like "a," "and," and "she"),[10] and *stem* their words, which removes the suffix of words (so "learns," "learning," and "learned" would all become "learn"). Collectively, these steps are known as *pre-processing* the text data. These steps vary based on the aim of the analysis, so should be carefully considered (Denny & Spirling, 2018).

---

[8] Some methods also account for word order in a circumscribed manner through using "bigrams" or "trigrams" as opposed to "unigrams." For example, the sentence "This class is hard" could produce unigrams ("this," "class," "is," "hard"), bigrams ("this class," "class is," "is hard,") or trigrams ("this class is," "class is hard").

[9] Popular packages in R for analyzing text data will do these for you. The tidytext and text mining ("tm") packages in R are particularly popular. Some analysis packages will also pre-process the data before conducting the analysis (e.g. the STM package in R) (Roberts, Stewart, & Tingley, 2015).

[10] These steps depend on the analysis being conducted. Researchers interested in linguistic style, for instance, may be primarily interested in function words instead of content words.

III. Applications of Text Analysis Techniques to an Experimental Study of Gender Bias in Education

A. Discussion Forum Experiment

In this section, we demonstrate both how to use lexical-based methods, supervised machine learning, and unsupervised machine learning and how to understand their affordances in the context of studying gender bias in an educational setting. Specifically, we rely on the text data from a field experiment conducted in the discussion board forums of 124 MOOCs. In this experiment, we created student profiles with names that sounded female or male, and randomized those fictitious students to post one of 32 different generic comments. We collected all of the written responses real instructors and students sent to these fictitious posters, which allows us to use automated text analysis to examine how the written responses differ when they are sent to the fictitious female students as opposed to the fictitious male students.

There are many reasons why we expect to see differences in how students and instructors respond to female and male students in online classes. Prior studies in face-to-face classes have shown that teachers encourage male students to ask more questions, make more comments and give more explanations than they do their female students (Crowley, Callanan, Tenenbaum, & Allen, 2001; M. G. Jones & Wheatley, 1990; S. M. Jones & Dindia, 2004; D. Sadker, Sadker, & Zittleman, 2009; M. Sadker & Sadker, 1984). Teachers and students both tend to overestimate male students' knowledge relative to female students' (Grunspan et al., 2016; Robinson-Cimpian, Lubienski, Ganley, & Copur-Gencturk, 2014). In online spaces, instructors are twice as likely to respond to white male MOOC participants than to other students (R. Baker, Dee, Evans, & John, 2018), participants in a mathematics online forum are less likely to give positive evaluations to females (Bohren, Imas, & Rosenberg, 2017), and participants in an economics job market forum are more likely to discuss personal information and physical appearance about women than about men (Wu, 2017). If females and males have different experiences in online classrooms, that may impact their course performance and persistence in the field, as females tend to

avoid majoring in subjects that have greater perceived gender biases (Ganley, George, Cimpian, & Makowski, 2017).

This experiment was conducted in 124 MOOCs from a major provider offered by 60 host institutions between August 1 and December 31, 2014.[11] The courses include many subjects, including chemistry, accounting, healthcare, neuroscience, film, and music. We posted eight comments in each of the courses, each with a race-gender evocative name for the following groups: white male, white female, black male, black female, Chinese male, Chinese female, Indian male, and Indian female.[12] The first and last name of each fictitious poster was displayed next to the comment in the discussion forum such that it was easily seen by every course participant who viewed the post in the discussion forum. The order of the eight comments was randomly chosen. To maximize statistical power, we combine the racial categories and compare responses to all male and all female fictitious posters. However, because a substantial number of MOOC participants were from the United States and may not be able to discern gender in Asian names, some of our analysis focuses only on the fictive students with white or black names.

In total, we posted 992 comments across these courses (eight comments placed in each of 124 courses). 798 of these comments received a response, and respondents were not significantly more likely to respond to the female or male randomized posters. We only analyzed the first response to each randomized post, as some of the later responses were responding to previous respondents (as opposed to the first poster). Only analyzing the first response allows us to be certain that the student or instructor was responding directly to the randomized poster. Six hundred twenty-one of the first respondents are students and 177 are instructors (which includes professors, teaching assistants, and staff members). In the

---

[11] We received IRB approval for this study, and we worked in close consultation with them to determine the number of comments placed in each course in order to minimize the costs placed on field participants.

[12] We used names that were recently used in studies that have experimentally manipulated perceptions of race and gender (e.g., Bertrand & Mullainathan, 2004; Milkman, Akinola, & Chugh, 2015; Oreopoulos, 2011). We chose a set of four first names and four last names for each gender-race combination (128 names in total).

following sections, we examine whether male and female students receive responses that are more positive or negative (using lexical-based methods), contain different levels of confirmation (using supervised machine learning), and contain different topics (using unsupervised machine learning).

B. Lexical-Based Methods

First, we test the hypothesis that female students may receive different amounts of positive or negative sentiment than male students, as prior studies have shown that teacher sentiment is associated with student performance (Brophy & Good, 1984). Because we are testing a specific hypothesis, we will use a supervised method. We have an idea of the types of words that should be included in positive versus negative posts, so we use lexical-based methods instead of supervised machine learning (although we could also choose to use supervised machine learning to answer this question) (see Figure 1). More specifically, we use two external sentiment dictionaries: Linguistic Inquiry and Word Count (LIWC) and Sentiment Analysis and Social Cognition Engine (SEANCE), then validate these measures in our sample. (See Section IIA for a fuller description of these two tools.)

LIWC measures sentiment through determining the percentage of words in each document that are found in the positive or negative dictionary (or word list). As shown in Table 1, 4.5 percent of words in respondents' posts are positive and 0.7 percent are negative. Instructors tend to use a higher percentage of positive words and a slightly lower percentage of negative words than students. Because SEANCE produces measures from a principal components analysis, the scale is not as easily interpretable as LIWC, but higher values also indicate more positive/negative text. SEANCE also shows that instructors use more positive words and fewer negative words than students.

Insert Table 1 here.

To validate the dictionary measures, we hand-code our data into positive, negative, and neutral comments to determine how the results from the dictionaries compare to human coding.[13] The hand-codes are considered to be the gold standard that the automated coding seeks to replicate. Table 2 shows the confusion matrices for LIWC and SEANCE, which allow us to directly compare the automated coding with the human coding.[14] Of the 241 messages that were hand coded as positive, LIWC categorized 208 as positive, 30 as neutral and 3 as negative. In comparison, SEANCE categorized 165 of the 241 as positive, 51 as neutral, and 25 as negative. To summarize this table for the positive codes, we estimate recall (the true positives out of the total hand-coded positives), precision (the true positives out of the computer's positives), accuracy (the percent of posts that the computer and human agree on), and the F-1 measure (a summary measure of recall and precision).[15] Table 3 (Panel a) shows these summary statistics. In our sample, LIWC categorizes a much higher percentage of positive posts as being positive (recall), but a higher percentage of the posts that SEANCE classified as positive were actually positive (precision). Despite LIWC's substantially higher recall score, both have similar F1-scores, because the F1-measure weights the lower performance metric more heavily. We also check whether misclassification varies by gender, and do not see evidence of this (see Panel b of Table 3).

---

[13] For comparison purposes, we convert the sentiment scores into positive/negative/neutral classifications. For LIWC, posts that have a higher percentage of positive terms than negative terms are classified as positive, posts that have a higher percentage of negative terms than positive terms are classified as negative, and posts that have an equal percentage of positive and negative terms are classified as neutral. For SEANCE, messages with a positive score above 0 and a negative score less than or equal to 0 are classified as positive, messages with a negative score above 0 and a positive score less than or equal to 0 are classified as negative, and all other posts are classified as neutral.

[14] See Section IIA for an introduction to the confusion matrix.

[15] The performance for the negative codes is worse, likely due to how rare the negative codes were (three percent of posts were negative).

Insert Table 2 here.

Insert Table 3 here.

The performance of neither of these measures is particularly strong, likely because neither was developed for our specific domain (i.e., MOOC discussion board fora). Because our dependent variable is binary, the measurement error in our dependent variable is non-classical and thus could cause some degree of bias in our treatment estimate. Because we have the true hand-coded variable, we can estimate this bias directly.[16]

Now that we have developed our measures, we can use our lexical-based measures as outcome variables in a traditional regression framework. Because this is a randomized experiment, these estimates give us a causal estimate of the effect of perceived poster gender on students' and instructors' language. We estimate the relationship between each outcome and the randomized poster's gender using the following model, which include fixed effects that parallel the randomization process:

$$Y_{ijkt} = \alpha + \beta Female_i + \theta_j + \delta_k + \mu_t + \epsilon_{ijkt}$$

where Y_{ijkt} represents the outcome measure used in response to randomized poster *i* in course *j* for comment *k* in the t^{th} position of the order of comments. *Female* is an indicator variable equal to 1 if the respondent is responding to a randomized poster with a female name, and 0 otherwise. θ_j represents course fixed effects, δ_k represents comment fixed effects, and μ_t represents comment order fixed effects. We show results both for all randomized posters, and for only the white and black randomized posters.

Table 4 shows the effects for the positive and negative sentiment analyses, using the LIWC, SEANCE, and hand-coded measures (for validation purposes). The analysis suggests that female posters are just as likely to receive a positive or negative post in response to their post.[17] Both LIWC and

[16] Meyer and Mitag (2017) also show how to estimate the degree of bias due to measurement error in binary dependent variables without having the true variable (i.e. the variable without measurement error).

[17] Recall that there are 241 positive posts and only 26 negative posts in our corpus (out of 798 posts).

SEANCE measures are within the confidence intervals of the hand-coded measures, which suggests minimal bias from measurement error (despite the relatively low performance of LIWC and SEANCE relative to the hand-coding).

Insert Table 4 here.

C. Supervised Machine Learning

Next, we would like to test whether female students receive different amounts of confirmation than male students. Teacher and student confirmation is fundamental to classroom communication, and prior research has found that confirmation can lead to higher levels of student learning, participation, motivation, connectedness, effort, and satisfaction (Campbell, Eichhorn, Basch, & Wolf, 2009; Ellis, 2000; Goodboy & Myers, 2008; LaBelle & Johnson, 2018; Sidelinger & Booth-Butterfield, 2010). We follow Johnson and Labelle's (2016) confirmation scale, which indicates the following four confirmation categories: assistance, acknowledgment, individual attention, and lack of disconfirmation. [18] In our context, some respondents may seek to provide concrete course assistance to the student ("assistance"), others may agree with and acknowledge the students' thoughts ("acknowledgment"), others may seek to connect with the randomized poster on an individual level ("individual attention"), and others may seek to undermine the students' thoughts ("disconfirmation").

Because this is another specific testable theory, we again use supervised methods. We do not have a clear sense of the words that might be included in the four categories, which means we would

[18] This scale is designed for student-to-student confirmation, and we apply it to both students and instructors. The instructor confirmation scale includes response to questions (which is very similar to the combination of assistance and acknowledgment, and includes answering students' questions fully and indicating that they appreciate student questions), demonstrating interest in the student (which is very similar to individual attention), teaching style (which does not apply in our online discussion forum context), and disconfirmation (which is included in the student-to-student confirmation scale) (Ellis, 2000).

struggle to apply lexical-based methods (see Figure 1). However, we can categorize our data by reading each document. Thus, we hand-code our data, then assess the performance of six different supervised machine learning algorithms intended to match the hand-coding. In a larger sample, we would use these algorithms to predict the observations that we do not hand code. However, we are able to hand code all of the data in our sample and we do so for exposition purposes.

We start with two coders manually coding each post into one or more of the four binary confirmation categories. The coders only saw the randomized comment and the respondent's comment when coding and were blind to the race and gender of the randomized poster and of the respondent. Any mentions of the randomized poster's name or the respondent's name were removed before coding. In most applications, we would only code a subset of responses due to resource constraints, but we code all responses for exposition purposes. Table 1 shows that assistance occurs in 72% of posts, acknowledgment in 21% of posts, individual attention in 24% of posts, and disconfirmation in 7% of posts. The inter-rater reliability for assistance is 86 percent, for acknowledgement is 91 percent, for individual attention is 83 percent, and for disconfirmation is 95 percent.

Table 5 shows the effects for the confirmation outcomes constructed using hand-coding. There are no statistically significant differences between female and male students' responses in the full sample. However, when we estimate the effects on white and black posters, respondents provide female students with 7 percentage points less assistance and 7 percentage points more individual attention than they do male students (Panel a).[19] As a qualified aside, we also explored these differences further by allowing these effects to differ by whether the respondent was a student or instructor. These exploratory analyses suggest that the differences in assistance levels are driven by instructors, who provide white and black females with 18 percentage points less assistance than they do white and black males. Because the

[19] We subset our sample to black and white posters because because a substantial number of MOOC participants were from the United States and may not be able to discern gender in Asian names.

identity of the first respondent is defined post-treatment, more research is needed to confirm these findings.

Insert Table 5 here.

We now assess how well supervised machine learning algorithms can mimic our hand-codes. We first prepare our data using the text mining ("tm") package in R to build a document-term matrix, which also allows us to convert all of our text to lower case, remove stopwords, stem our terms, remove all punctuation, and remove words that occur in fewer than 1% of the posts. We then prepare our data for 10-fold cross-validation by randomly partitioning our data into 10 subsets. We iteratively train our models on 90% of the data, then predict the codes for the remaining 10% of the data. By repeating this procedure 10 times, we have a prediction for each observation.[20] We compare six different models: ordinary least squares (OLS), LASSO, ridge, elastic net, support vector machine (SVM) and random forests. We can assess the performance of each model by comparing the hand-codes to the predicted codes from cross-validation and choose the model that performs the best using our metrics of precision, recall, and the F1-measure.

The low prevalence of acknowledgment, individual attention, and disconfirmation makes it difficult for the machine learning algorithm to predict these categories. Assistance occurs in 576 responses, but acknowledgement only occurs in 170 responses, individual attention in 190 responses, and disconfirmation in 59 responses. Table 6 shows the performance of the models used to predict assistance. LASSO, ridge, and elastic net are the most precise, and SVM and random forest have higher recall. Overall, LASSO, ridge, elastic net, and random forest have the best performance (as determined by the F1-measure), and we use these in our analyses. In contrast, for acknowledgment the F1-measure varies between 0.23 and 0.48, for individual attention it varies between 0.16 and 0.40, and for disconfirmation it

[20] See Section IIB for an introduction to k-fold cross-validation.

varies between 0.06 and 0.10.[21] Only the performance of assistance is high enough to reasonably use in analysis, so we focus on this measure.

Insert Table 6 here.

Table 7 shows the effects for assistance using supervised machine learning. In the main sample, the predicted estimates are all within the confidence interval of the hand-coded measure of -0.06 to 0.03. However, when we subset the sample to the 408 observations in which the randomized poster was black or white, 3 of the 4 predicted estimates are outside of the confidence interval of the hand-coded measure (of -0.14 to 0.00). This suggests that we would need to achieve a higher predictive performance to be able to conduct estimates on our subsamples to avoid substantial measurement error bias.[22]

Insert Table 7 here.

D. Unsupervised Machine Learning

We are also interested in whether respondents may discuss different topics in response to male and female students. Because we do not have specific topics in mind that we would like to test, using an unsupervised method is the most appropriate choice (see Figure 1). This allows our text data to form its own groupings (based off of the number of topics we choose), and we subsequently observe whether

[21] Disconfirmation may also be a more complex task to predict. Disconfirmation sometimes includes clear terms of disagreement, like "no" and "not really," but often is more complex. For instance, one disconfirming post simply counters "highly subjective" when a fictitious poster complained about the lectures, and another states that it is the "calm before the storm" when a fictitious poster stated that they were feeling confident about the course. None of these terms (subjective, calm, storm) shows up in any of the other disconfirming posts.

[22] As noted earlier, this is primarily an issue with binary variables, which do not exhibit classical measurement error.

these groupings vary by the gender of the randomized poster. To allow the topics to vary by the fictitious students' gender, we will use a structural topic model.[23]

To estimate our structural topic model, we use the STM package in R (Roberts, Stewart, & Tingley, 2018). We first preprocess the text data by stemming all the words (including all words in their root form), removing punctuation, and removing stop words (function words). We specify 20 topics and model the topic prevalence as a function of gender (our treatment of interest) and the 32 seed comments.

The model produces twenty clusters, which we then label. First, we examine the words with the highest probabilities and highest FREX and label the topics accordingly. Table 8 shows these (stemmed) words for each topic. For example, the highest probability words for the first topic are "video," "watch," and "minute," and we name this topic "watching videos." We repeat this process to label all twenty topics.

Insert Table 8 here.

Figure 2 shows the results from the topic models. This figure tells us that the topic "assignments" comprise 7 percent of the topics discussed with male posters, and 6 percent of the topics discussed with female posters. Only two out of the twenty topics discussed statistically significantly differ between male and female students. "Concepts" is discussed more with female students and "Understanding Topics" is discussed more with male students. Concepts contains words like "new," "change," and "concept," and Understanding Topics contains words like "yes," "understand," and "think" (see Table 8).

Insert Figure 2 here.

As mentioned in Section IIC, topic modeling is primarily useful for data exploration and theory generation. If we were further interested in understanding the gender differences around discussions of concepts and understanding topics, we could construct these measures using supervised machine learning methods to test whether this relationship holds. Using supervised machine learning would allow us to

[23] See Section IIC for an introduction to topic models.

explicitly test how the automated measure compares to our conceptualization of the measures using hand-coding as the gold standard, which is harder to do with unsupervised machine learning.

IV. Discussion and Conclusion

In this study, we provided a comparative overview of a variety of techniques that have recently emerged for the study of text as data in the social sciences, including lexical-based methods, supervised machine learning, and unsupervised machine learning. We illustrated the use of these techniques (and their comparative properties) through the empirical example of gender bias in online classrooms using data from a field experiment in which the identities of discussion forum posters were randomly manipulated. We show that respondents provide 7 percentage points less assistance and 7 percentage points more individual attention to (the black and white) female posters. We see no significant differences in positive and negative sentiment, and small differences in topics discussed in response to male and female posters.

Each of these methods has its promises and pitfalls. Lexical-based methods are generally faster and easier to use than supervised machine learning methods but can perform poorly when applied to a different sample or domain and require validation. In our empirical application, we showed that less than half of the instances that our two lexical-based methods predicted as being positive were actually positive when compared to human coding. However, the relatively poor prediction performance did not induce much bias in our experimental analysis. Supervised machine learning requires a higher start-up cost of hand-coding, but it can perform well when it is trained in a similar sample with large enough samples. In our empirical application, we showed that our supervised machine learning measure performed reasonably well for sufficiently prevalent codes and when tested in sufficiently large samples (i.e. not small subsamples). Unsupervised methods are more appropriate for theory generation as opposed to the theory testing abilities of the first two methods, and we demonstrated how structural topic models can be used to examine how topic clusters can vary by treatment status or other document characteristics.

The performance of automated text methods is constantly improving. Natural language processing tools now allow researchers to identify parts of speech, named entities (like people and places), parsing sentences into their grammatical structures, and recognizing co-references between noun phrases, and determining the syntactic structure of sentences (Hirschberg & Manning, 2015). Researchers are also increasingly taking word similarities into account in their analyses by representing words as vectors, in which words with similar meanings (e.g., “course” and “class”) have similar vector representations (known as word embeddings) (Mikolov, Chen, Corrado, & Dean, 2013). Deep learning methods (such as neural networks) are also beginning to be implemented in educational settings, and can serve as a substantial improvement over more traditional supervised machine learning methods (Khajah, Lindsey, & Mozer, 2016). As these methods improve, the research questions we can answer and the data sets we can analyze continue to expand.

References

Anderson, R. (2018). Beyond copy room collaboration: A case study of online informal teacher professional learning. In *Rethinking Learning in the Digital Age: Making the Learning Sciences Count, 13th International Conference of the Learning Sciences (ICLS)* (Vol. 3, pp. 1511–1512). London, UK: International Society of the Learning Sciences.

Baker, R., Dee, T., Evans, B., & John, J. (2018). *Bias in Online Classes: Evidence from a Field Experiment* (CEPA Working Paper No. 18–03). Retrieved from http://cepa.stanford.edu/wp18-03

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636.

Beattie, G., Laliberté, J. W. P., & Oreopoulos, P. (2018). Thrivers and divers: Using non-academic measures to predict college success and failure. *Economics of Education Review*, *62*(January 2017), 170–182.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, *94*(4), 991–1013.

Bettinger, E., Liu, J., & Loeb, S. (2016). Connections Matter: How Interactive Peers Affect Students in Online College Courses. *Journal of Policy Analysis and Management*. https://doi.org/10.1002/pam

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bohren, J. A., Imas, A., & Rosenberg, M. (2017). *The Dynamics of Discrimination: Theory and Evidence* (PIER No. 17–021).

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Brophy, J., & Good, T. L. (1984). Teacher behaviour and student achievement. In *Handbook of research on teaching* (3rd ed., pp. 328–375).

Burstein, J., Sabatini, J., & Shore, J. (2014). Natural Language Processing for Educational Applications. In *The Oxford Handbook of Computational Linguistics* (2nd ed.).

Campbell, L. C., Eichhorn, K. C., Basch, C., & Wolf, R. (2009). Exploring the relationship between teacher confirmation, gender, student effort in the college classroom. *Human Communication*, *12*(4), 447–464.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, *49*(3), 803–821.

Crowley, K., Callanan, M. A., Tenenbaum, H. R., & Allen, E. (2001). Parents Explain More Often To Boys Than To Girls During Shared Scientific Thinking, *12*(3), 258–261.

Denny, M., & Spirling, A. (2018). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. *Political Analysis*. https://doi.org/10.2139/ssrn.2849145

Ellis, K. (2000). Perceived Teacher Confirmation. *Human Communication Research*, *26*(2), 264–291.

Evans, B. J., Marsicano, C. R., & Lennartz, C. J. (2019). Cracks in the Bedrock of American Democracy: Differences in Civic Engagement Across Institutions of Higher Education. *Educational Researcher*, *48*(1), 31–44.

Fesler, L. (2019). Understanding How Virtual College Counselors Can Nudge Students Through the College Application Process. In *Society for Research on Educational Effectiveness*.

Ganley, C. M., George, C. E., Cimpian, J. R., & Makowski, M. B. (2017). Gender Equity in College Majors: Looking Beyond the STEM/Non-STEM Dichotomy for Answers Regarding Female Participation. *American Educational Research Journal*, 1–35.

Gegenheimer, K., Hunter, S., Koedel, C., & Springer, M. (2018). Does Feedback Matter? Performance Management and Improvement in Public Education.

Gentzkow, M., Kelly, B. T., & Taddy, M. (2017). *Text as Data* (NBER Working Paper Series No. 23276). Cambridge, MA.

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and Combining Sentiment Analysis Methods. In *Proceedings of the first ACM conference on Online social networks* (pp. 27–37). Boston, MA: ACM.

Goodboy, A. K., & Myers, S. A. (2008). The Effect of Teacher Confirmation on Student Communication and Learning Outcomes. *Communication Education*, *47*(2), 153–179.

Graesser, A. C., Mcnamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, *40*(5), 223–234.

Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, *108*(7), 2643–2650.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Grunspan, D. Z., Eddy, S. L., Brownell, S. E., Wiggins, B. L., Crowe, A. J., & Goodreau, S. M. (2016). Males under-estimate academic performance of their female peers in undergraduate biology classrooms. *PLoS ONE*, *11*(2), 1–16.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.

Hirschberg, J., & Manning, C. D. (2015). Advances in Natural Language Processing. *Science*, *349*(6245).

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247.

Jerzak, C. T., King, G., & Strezhnev, A. (2018). *An Improved Method of Automated Nonparametric Content Analysis for Social Science*.

Johnson, Z. D., & LaBelle, S. (2016). Student-to-Student Confirmation in the College Classroom: An Initial Investigation of the Dimensions and Outcomes of Students' Confirming Messages. *Communication Education*, *65*(1), 44–63.

Jones, M. G., & Wheatley, J. (1990). Gender differences in teacher-student interactions in science classrooms. *Journal of Research in Science Teaching*, *27*(9), 861–874.

Jones, S. M., & Dindia, K. (2004). A Meta-Analytic Perspective on Sex Equity in the Classroom. *Review of Educational Research*, *74*(4), 443–471.

Jurafsky, D., & Martin, J. H. (2018). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed.).

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. (2018). Automatically Measuring Question Authenticity in Real-World Classrooms. *Educational Researcher*.

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How deep is knowledge tracing? In *EDM 2016*.

LaBelle, S., & Johnson, Z. D. (2018). Student-to-student confirmation in the college classroom: the development and validation of the Student-to-Student Confirmation Scale. *Communication Education*, *67*(2), 185–205.

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35–65.

Meyer, B. D., & Mittag, N. (2017). Misclassification in binary choice models. *Journal of Econometrics*, *200*(2), 295–311. https://doi.org/10.1016/j.jeconom.2017.06.012

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *arXiv preprint arXiv:1301.3781* (pp. 1–12).

Milkman, K. L., Akinola, M., & Chugh, D. (2015). What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway into Organizations. *Journal of Applied Psychology*, *100*(6), 1678–1712.

Mockler, N. (2018). Discourses of teacher quality in the Australian print media 2014–2017: a corpus-assisted analysis. *Discourse: Studies in the Cultural Politics of Education*, (December).

Nelson, L. K. (2016). Computational Grounded Theory: A Methdological Framework. *Sociological Methods & Research*. https://doi.org/10.1177/0049124117729703

Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy*, *3*(4), 148–171.

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of ACL*.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*.

Penner, E. K., Rochmes, J., Liu, J., Solanki, S., & Loeb, S. (2019). Differing Views of Equity: How

Prospective Teachers Perceive Their Role in Closing Achievement Gaps. *The Russell Sage Foundation Journal of the Social Sciences*.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.

Reich, J., Stewart, B., Mavon, K., & Tingley, D. (2016). The civic mission of MOOCs: Measuring engagement across political differences in forums. *L@S 2016 - Proceedings of the 3rd 2016 ACM Conference on Learning at Scale*, 1–10. https://doi.org/10.1145/2876034.2876045

Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. E., & Stewart, B. M. (2015). Computer-Assisted Reading and Discovery for Student-Generated Text in Massive Open Online Courses. *Journal of Learning Analytics*, *2*(1), 156–184.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2015). stm: R Package for Structural Topic Models. *Journal of Statistical Software*, *VV*(2014).

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016). Navigating the Local Modes of Big Data. In R. M. Alvarez (Ed.), *Computational Social Science: Discovery and Prediction*. Cambridge University Press.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2018). stm: R Package for Structural Topic Models. *Journal Of Statistical Software*.

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology*, *50*(4), 1262–1281.

Sadker, D., Sadker, M., & Zittleman, K. R. (2009). *Still failing at fairness: How gender bias cheats girls and boys and what we can do about it*. New York, NY: Charles Scribner.

Sadker, M., & Sadker, D. (1984). *Promoting Effectiveness in Classroom Instruction. Year 3: Final*

Report. Andover, MA.

Sajjadiani, S., Sojourner, A., Kammeyer-mueller, J., & Mykerezi, E. (2019). Using Machine Learning to Translate Applicant Work History into Predictors of Performance & Turnover.

Sidelinger, R. J., & Booth-Butterfield, M. (2010). Co-constructing student involvement: An examination of teacher confirmation and student-to-student connectedness in the college classroom. *Communication Education*, *59*(2), 165–184.

Stone, P., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Cammbridge, MA: MIT Press.

Wu, A. H. (2017). *Gender Stereotyping in Academia : Evidence from Economics Job Market Rumors Forum*. University of California, Berkeley.

Table 1: Descriptive Statistics

Outcome	All	Student	Instructor
Positive (LIWC) (%)	4.5	4.2	5.8
Positive (SEANCE) (PC)	0.07	0.02	0.26
Negative (LIWC) (%)	0.7	0.8	0.6
Negative (SEANCE) (PC)	-0.39	-0.34	-0.57
Assistance(%)	72.2	69.2	82.5
Acknowledgment(%)	21.3	20.5	24.3
Individual Attention (%)	23.8	23.5	24.9
Disconfirmation(%)	7.4	9.2	1.1
N	798	621	177

The instructor category includes professors, TAs, and staff.

This table reads: 72 percent of randomized posters who received a response received assistance.

69 percent of randomized posters who received a response from a student received assistance.

83 percent of randomized posters who received a response from a student received assistance.

The SEANCE measures are principal components (PCs).

Table 2: Confusion Matrices for Lexical-Based Methods (LIWC and SEANCE)

(a) LIWC

Predicted	Hand Coding			
	Positive	Neutral	Negative	Total
Positive	208	253	10	471
Neutral	30	239	8	277
Negative	3	39	8	50
Total	241	531	26	798

(b) SEANCE

Predicted	Hand Coding			
	Positive	Neutral	Negative	Total
Positive	165	166	3	334
Neutral	51	253	10	314
Negative	25	112	13	150
Total	241	531	26	798

Confusion matrices allow us to validate the performance of LIWC and SEANCE in our sample, by determining how frequently LIWC and SEANCE agree with our hand codes. Each post was hand-coded as positive, negative or neutral, and we also convert the LIWC and SEANCE measures from continuous measures into positive, negative, or neutral. This table reads: of the 241 posts that were positive (according to our hand coding), 208 were positive according to LIWC and 165 were positive according to SEANCE.

Table 3: Prediction Performance of Lexical-Based Methods (LIWC and SEANCE) for Positive Codes

(a) All

Measures	LIWC	SEANCE
Recall	0.86	0.68
Precision	0.44	0.49
Accuracy	0.57	0.54
F1	0.58	0.57

(b) By Gender

Measures	LIWC		SEANCE	
	Male	Female	Male	Female
Recall	0.85	0.88	0.69	0.68
Precision	0.43	0.45	0.49	0.49
Accuracy	0.56	0.58	0.53	0.55
F1	0.57	0.60	0.58	0.57

This table summarizes the predictive performance of LIWC and SEANCE. Recall measures the percentage of the hand coded positive instances that were predicted by the computer as being positive, precision measures the percentage of the computer's positive predictions that were hand-coded as being positive, accuracy measures the percentage of the computer's predictions that were correct out of all posts, and the F1-measure is the harmonic mean of precision and recall. All of these statistics can be calculated directly from Table 2.

Table 4: Estimated Effect of Perceived Gender on Positive and Negative Sentiment

(a) Positive

Sample	LIWC	SEANCE	Hand-Coded	N
All	0.003	-0.000	0.004	798
	(0.034)	(0.033)	(0.028)	
White and Black	-0.077	0.003	-0.013	408
	(0.049)	(0.050)	(0.042)	

(b) Negative

Sample	LIWC	SEANCE	Hand-Coded	N
All	0.006	-0.010	-0.013	798
	(0.018)	(0.025)	(0.013)	
White and Black	0.026	-0.010	0.001	408
	(0.026)	(0.035)	(0.020)	

Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** p<.01
Includes course, comment, and sequence fixed effects.
Standard errors are clustered at the course level.

Table 5: Estimated Effect of Perceived Gender on Student and Instructor Confirmation

Sample	Assistance	Acknowledgment	Indiv Att	Disconfirmation	N
All	-0.017	0.018	0.021	0.023	798
	(0.024)	(0.022)	(0.031)	(0.019)	
White and Black	-0.069*	0.040	0.070+	0.037	408
	(0.034)	(0.034)	(0.041)	(0.029)	

Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** p<.01
Includes course, comment, and sequence fixed effects.
Standard errors are clustered at the course level.

Table 6: Prediction Performance of Supervised Machine Learning Models: Assistance

	OLS	LASSO	Ridge	Elastic-Net	SVM	Random Forest
Recall	0.77	0.76	0.75	0.75	0.81	0.81
Precision	0.73	0.97	0.98	0.97	0.72	0.82
Accuracy	0.65	0.75	0.75	0.75	0.68	0.73
F1	0.75	0.85	0.85	0.85	0.76	0.82

This table summarizes the predictive performance of six different methods in predicting the assistance code. Recall measures the percentage of the hand coded positive instances that were predicted by the computer as being positive, precision measures the percentage of the computer's positive predictions that were hand-coded as being positive, accuracy measures the percentage of the computer's predictions that were correct out of all posts, and the F1-measure is the harmonic mean of precision and recall.

Table 7: Estimated Effect of Perceived Gender on Course Assistance: Hand-Coded vs. Predicted

Sample	Hand-Coded	LASSO	Ridge	Elastic Net	Random Forest	N
All	-0.017	0.022	0.008	0.026	0.030	798
	(0.024)	(0.017)	(0.018)	(0.018)	(0.030)	
White and Black	-0.069	-0.008	0.020	0.003	0.045	408
	(0.034)	(0.023)	(0.024)	(0.023)	(0.046)	

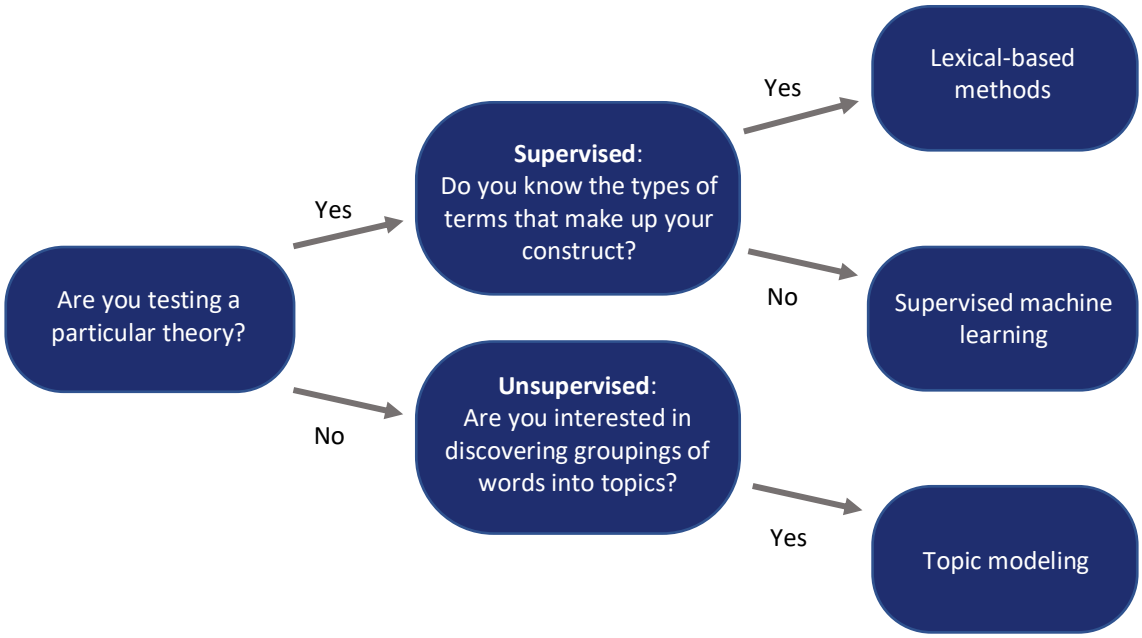
Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** p<.01
Includes course, comment, and sequence fixed effects.
Standard errors are clustered at the course level.

Table 8: Topics in Discussion Forum Responses

Label	Stems
Watching Videos	video, watch, minut
Completing	cours, grade, complet
Concepts	new, cours, chang, concept
Materials	lecture, learn, android, pdf
Help	help, can, find
Understand Topics	yes, understand, think
Practice	practic, use, class, data
Learning	learn, video, edit
Resources	course, take, resources
Work Strategies	start, unit, procrastin
Assignments	assign, quiz, submit
Resume	resum, follow, job
Prior Background	concept, background, problem
Liking Course	like, course, enjoy, agree
Readings	week, lectur, read, watch
Research	research, product, work
Program	program, game, course
Essay	read, essay, behind
Unenrolling	course, unenrol, click
Projects	project, practic, import

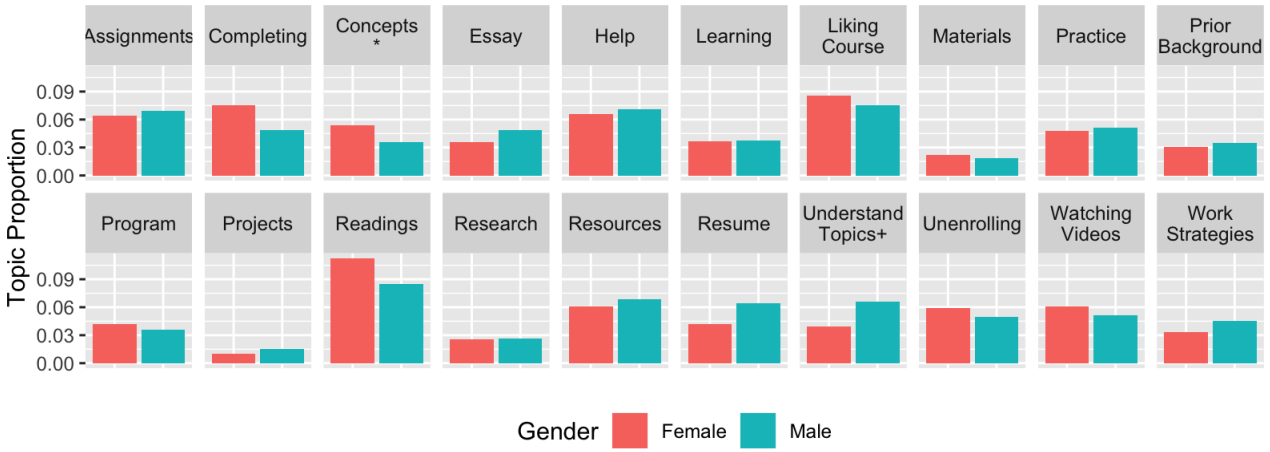
This shows the twenty topics that emerged from the structural topic model. The "stems" column shows a subset of the words that are the most probable or have the highest FREX (which captures the frequency and exclusivity of given stems) (Roberts, Stewart & Tingley, 2017). The labels were determined based on the stem words for each topic.

Figure 1: Choosing the Appropriate Text-as-Data Method



Note: these categories are not mutually exclusive nor exhaustive. In particular, lexical-based methods can lead to an outcome of interest (as it is in this chart) or be used as an input in supervised machine learning. In addition, there are many types of unsupervised machine learning that we do not cover here due to space limitations (although topic models are one of the most commonly used methods).

Figure 2: Estimated Effect of Perceived Gender on Respondent Topics



$+\ p < 0.10$, * $p < 0.05$, ** p<.01

Includes comment fixed effects.