Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–June: 1–17

DOI: 10.1177/20539517221149106
journals.sagepub.com/home/bds
SAGE

Formally comparing topic models and human-generated qualitative coding of physician mothers' experiences of workplace discrimination

Adam S Miner[1,2,†], Sheridan A Stewart[3,†], Meghan C Halley[4], Laura K Nelson[5] and Eleni Linos[2,6]

Abstract

Differences between computationally generated and human-generated themes in unstructured text are important to understand yet difficult to assess formally. In this study, we bridge these approaches through two contributions. First, we formally compare a primarily computational approach, topic modeling, to a primarily human-driven approach, qualitative thematic coding, in an impactful context: physician mothers' experience of workplace discrimination. Second, we compare our chosen topic model to a principled alternative topic model to make explicit study design decisions meriting consideration in future research. By formally contrasting computationally generated (i.e. topic modeling) and human-generated (i.e. thematic coding) knowledge, we shed light on issues of interest to several audiences, notably computational social scientists who wish to understand study design tradeoffs, and qualitative researchers who may wish to leverage computational methods to improve the speed and reproducibility of labor-intensive coding. Although useful in other domains, we highlight the value of fast, reproducible methods to better understand experiences of workplace discrimination.

Keywords

Attitude of health personnel, topic model, natural language processing, qualitative research, occupational health, motherhood penalty

Introduction

Early work validating unsupervised computational text analysis methods, particularly topic modeling, involved humans interpreting and validating computationally generated themes (Chang et al., 2009; Grimmer and Stewart, 2013; Mimno et al., 2011). This approach helped to validate the use of computational text analysis, although researchers have continued to rely on automated methods of evaluation (Röder et al., 2015). However, these methods vary considerably and may be less correlated with human judgments than assumed (Hoyle et al., 2021). Few studies have evaluated whether human coders would have generated the same themes as unsupervised methods (Nelson et al., 2021; cf. Baumer et al., 2017) or whether the latter methods could help to surface themes missed by human coders.

This study formally compares computationally generated topics to human-generated codes using a dataset of clear social and medical import: physician mothers' descriptions of workplace discrimination (Halley et al., 2018). This exercise provides a novel opportunity to assess how closely standard computational methods might reflect rigorous human coding and analysis with the aim of informing big data practices.

[1]Department of Psychiatry and Behavioral Sciences, Stanford University, Palo Alto, California, USA
[2]Department of Epidemiology and Population Health, Stanford University, Palo Alto, California, USA
[3]Department of Sociology, Stanford University, Stanford, California, USA
[4]Center for Biomedical Ethics, Stanford University, Stanford, California, USA
[5]Department of Sociology, University of British Columbia, Vancouver, British Columbia, Canada
[6]Department of Dermatology, Stanford University, Stanford, California, USA

[†]These authors contributed equally to this manuscript.

Corresponding author:
Sheridan A Stewart, Department of Sociology, Stanford University, Stanford, California, United States.
Email: sastew@stanford.edu

Topic modeling is an established computational approach to summarizing a corpus of text documents (Blei, 2012). Within the health domain, topic modeling has been utilized to summarize documents without *a priori* human review in mental health, electronic medical record review, and medically relevant social media posts (Arnold et al., 2016; Gaut et al., 2017; Lehman et al., 2012; Lossio-Ventura et al., 2021; Paul and Dredze, 2014; Schweinberger et al., 2021).

While computational social science has leveraged the democratization of computing power, data, and programming skills, it remains unclear how well approaches like topic modeling approximate human expert-guided analysis of the same documents or how useful computational methods can be for identifying themes missed by human experts. Additionally, computationally derived topics may or may not be similar to the kinds of concepts social scientists analyze (Ylä-Anttila et al., 2022). Explicit writing about the decisions and tradeoffs is rare (with notable exceptions, e.g. Nelson, 2020; Nelson et al., 2021; Roberts et al., 2014), which impedes such comparisons. Because policymakers, funding agencies, and other powerful actors have begun to embrace computational methods, asymmetries of understanding can disadvantage researchers who cannot evaluate when a computational approach could be used to study new questions or to study old questions in a new light.

Researchers have developed numerous approaches to topic modeling (Blei and Lafferty, 2006, 2007; Gerlach et al., 2018; Ramage et al., 2009; Ramage et al., 2011; Roberts et al., 2014; Zhao et al., 2021), with no consensus as to which approach is appropriate for a given context. We chose a well-established approach to topic modeling, latent Dirichlet allocation (LDA; Blei et al., 2003), for two reasons: 1) LDA has been widely used and thus facilitates comparisons to prior research, and 2) LDA is used by sophisticated computational social scientists while remaining accessible to researchers with less exposure to computational methods.

By leveraging the results of a well-designed, previously published qualitative study (described below), we compare how humans and LDA extract and summarize meaning from the same text. Through formal evaluation and by publishing our code, we aim to surface epistemological issues central to knowledge production in both social science and medicine. That is, we formalize and make explicit differences in how computers and people "read" text. This work may increase advocacy for groups experiencing discrimination by facilitating faster and more reproducible research on these issues.

Case study

In academic medicine, women receive lower pay and are less likely to reach the rank of full professor (Jena et al., 2015; Jena et al., 2016). Discrimination regarding childbearing and child-rearing has been proposed as one mechanism by which physician mothers are disadvantaged throughout their careers (Adesoye et al., 2017). Indeed, there are long-standing trends in workplace disadvantage for mothers across workplace settings and countries (Aisenbrey et al., 2009; Correll et al., 2007; Yu and Kuo, 2017). Although gathering and summarizing experiences of discrimination are often labor-intensive, online surveys have facilitated larger, faster, and more widely distributed data collection on physician mothers' lived experiences (Halley et al., 2018; Melville et al., 2019).

A study of nearly 6000 physician mothers indicated 73% experienced workplace discrimination based on gender or motherhood. This survey also included an open-ended question asking respondents to describe their experiences of discrimination (Adesoye et al., 2017). To analyze the 1,009 responses generated from this question, experienced qualitative researchers utilized an inductive, iterative approach to develop a set of codes representing key topics in the data (Halley et al., 2018). The coded data were then used for additional analysis, using a grounded theory approach, to examine relationships among themes and develop a conceptual framework illustrating these relationships. Experiences of discrimination are unique but also follow trends that, when consolidated and described, are crucial to remediate context-specific discrimination (Halley et al., 2018).

Human-generated coding (e.g. as part of qualitative thematic analysis) is a well-established approach to inductively identifying common ideas across individual survey responses. While this type of detailed analysis provides unique insights into experiences of discrimination among female physicians and generates hypotheses for future research, the methods also have significant limitations. One limitation is the time and effort required to identify relevant codes based on latent topics within the qualitative data and then refine and apply them with high accuracy and consistency across human coders. The effort required limits the extent to which qualitative analysis can be applied to sources of large-N qualitative data like large or repeated surveys with open-ended questions. Computational approaches may be able to augment the capacity of human-generated analysis of large bodies of qualitative data.

In contrast to prior work, we do not use supervised machine learning to evaluate large corpora based on qualitative analysis of samples of the texts because we seek to compare both approaches on an equal footing. Neither approach uses prior knowledge of latent structure. Supervised machine learning requires a qualitative analysis upfront (i.e. generating and applying labels), which may miss themes that might otherwise be identified inductively. We use an unsupervised approach, which we compare to the ground truth of the qualitative thematic analysis. By "ground truth," we do not mean the qualitative analysis has produced a complete and objective model of this

corpus or workplace discrimination. Rather, we view the qualitative analysis as a rigorous baseline that is useful for comparison. This comparison may indicate how well a computational approach can reflect the results of more time-intensive qualitative analysis and, conversely, how far computational approaches can get researchers in jumpstarting qualitative analyses.

The present study addresses the following research questions:

- Research Question 1: What types of discrimination are reported by physician mothers?
- Research Question 2: Does topic modeling with LDA surface the same ideas as human-generated thematic analysis?
- Research Question 3: Is the comparability of computer-generated topics to human-generated codes robust to the number of topics the researcher chooses?

Data and methods

Data

Our data were collected during a survey of members of the Physician Moms Group, a Facebook group for physician mothers. This 2016 survey concerned health and experiences of discrimination. At the time, the group had approximately 35,000 active members. Prior quantitative research (Adesoye et al., 2017) reports the results from the survey ($N = 5782$) and provides detailed information about the procedure and sample.

Two authors (MCH and EL) and colleagues previously conducted a qualitative analysis of free-text responses to an open-ended question included in the survey (Halley et al., 2018). Specifically, after responding to several items about experiences of discrimination, respondents read the prompt, "We want to hear your story and experience. Please share."

This qualitative analysis is the baseline to which we compare our computational approach. For this analysis, multiple analysts first read the responses to the open-ended question in detail and generated a structured codebook of concepts or ideas present in the data. They then went through a rigorous and time-intensive process of refining the codebook to clarify each code's meaning, including multiple rounds of formal inter-rater reliability assessment. Once they had established excellent inter-rater reliability, they reviewed the dataset again, applying all relevant codes. We use the original 17 codes from this qualitative analysis to evaluate the results of our computational approach, as they represent the closest analog to the topics generated through computational approaches and were systematically applied to the entire dataset.

In the prior analysis, 947 of the 1,009 free-text responses were deemed relevant to workplace discrimination. In our computational analysis, we imposed no such constraint. After preprocessing, our corpus consists of 988 documents.

Preprocessing

Researchers using computational tools like LDA to study meaning often assume there is a correspondence between the meaning of a word and the contexts in which it is used (Sahlgren, 2008). Synonyms present the extreme case: they can be substituted for one another in many utterances without appreciably changing the utterances' intended or received meaning. According to this assumption, the similarity of two words is proportional to the extent to which they are used in similar contexts—that is, the extent to which they co-occur with the same *context words*. Notably, this assumption can also treat antonyms as similar due to their occurrence in similar contexts. LDA builds on this assumption by assuming the existence of latent themes comprising sets of words that frequently co-occur.

Algorithms like LDA cannot "read" like a human can because they do not know what words mean (Rhody, 2012). The sets of words identified by LDA—called topics—are sensitive to the exact form of each word. Capitalization, punctuation, and conjugation affect the patterns the algorithm identifies. Researchers' choices about cleaning the data thus affect how the algorithm works (Denny and Spirling, 2018; Krouska et al., 2016; Schofield and Mimno, 2016). These decisions relate to assumptions about language (such as the distributional hypothesis) that researchers rarely describe when reporting their methods.

Because LDA learns about words and identifies topics based on word co-occurrence without knowledge of words' meanings, we wanted to ensure that our models treated different versions of the same word as one. LDA on its own cannot distinguish between tokens like "Physician," "physicians," and "physician." (including the period) like a human can. Instead, LDA will treat each as a different word, or *type*. Treating different forms of the same word as distinct can result in considerable noise and, potentially, worse topics. A large vocabulary (unique words in the corpus) also makes training models more computationally demanding. Therefore, we undertook several steps to convert variations of a word to one type during preprocessing, like "physician" in the example above. We describe preprocessing steps explicitly to bridge understanding between computational and qualitative readers and elucidate ways that humans and machines may "read" the same text differently.

The first row of Table 1 presents a sentence as it originally appeared in a survey response. The word "Told" is capitalized. Lowercasing the corpus forces LDA to treat all occurrences of "Told" and "told" as the same type. We also expanded contractions ("shouldn't" → "should

not") and standardized ordinal numbers (e.g. "1st" → "first"), reducing the vocabulary size and ensuring LDA treats variants as the same type. These changes are part of the process of *normalizing* the text, eliminating stylistic differences that may otherwise obscure latent themes. For example, some writers may write more formally, even when talking about the same themes. Stylistic differences can be valuable to study, but for our research questions, we must minimize these differences and focus on latent themes.

We also removed stop words—for example, "the" or "and"—that provide little information about the meaning of neighboring words due to their prevalence (Gerlach et al., 2019). Schofield and colleagues (2017) note that models may only be affected by removing the most frequent stop words. We adopted the widespread practice of removing stop words from the start using popular lists from the NLTK (Bird et al., 2009) and spaCy (Honnibal et al., 2020) libraries for Python. We also removed parts of speech such as punctuation and numbers using part-of-speech tagging with spaCy. Although a number can be informative—for example, in a mention of a salary or pay gap—LDA will treat each unique number as a different type. Unless a number is repeated many times, it is unlikely LDA will associate it with latent themes in a useful way.

When working with text, researchers also often choose to either lemmatize or stem the corpus. *Lemmatization* converts variations of a word to a root word (e.g. "discrimination" → "discriminate"), which helps with reducing the vocabulary size and forcing LDA to treat variations of a word as the same type. *Stemming* has a similar goal and can be more computationally efficient, but it can be cruder: stemming removes the ends of words (e.g. "discrimination" → "discrimin") and sometimes fails to convert variations to the same root. Notably, more aggressive approaches to stemming have been found to lead to worse models according to some criteria (Schofield and Mimno, 2016). We lemmatized the documents using spaCy (Honnibal et al., 2020).

Finally, we identified n-grams (sequences of length n) using the Gensim library (Řehůřek and Sojka, 2010). Before n-grams are identified, researchers determine the lengths to consider (one word, two words, and so on), the minimum frequency, and sometimes other parameters. We allow for bigrams ($n = 2$) and trigrams ($n = 3$), replacing the whitespace in n-grams with underscores. Each bigram or trigram thus becomes a single token. This is another stage at which researchers' decisions matter. Given our priors about the data, had "sick leave" and "maternity leave" not been identified as bigrams ("sick_leave" and "maternity_leave," respectively), we may have repeated the process with different parameters.

Table 1. Effects of preprocessing steps.

Step	Resulting Text (Changes Bolded)
Original Text	Told "pregnancy was a choice… not an illness" so shouldn't be allowed to use sick leave for maternity leave.
Lowercasing	**told** "pregnancy was a choice… not an illness" so shouldn't be allowed to use sick leave for maternity leave.
Expanding contractions and ordinal numbers	told "pregnancy was a choice… not an illness" so **should not** be allowed to use sick leave for maternity leave.
Lemmatizing, removing stop words, and removing unneeded parts of speech	**tell** pregnancy choice illness **allow** use sick leave maternity leave.
Identifying bigrams and trigrams	tell pregnancy choice illness allow use **sick_leave maternity_leave**.

Specifying the topic model

Researchers using topic modeling algorithms face two significant challenges. The first challenge is selecting the number of topics. LDA itself is designed to populate a certain number of topics (k), which must be prespecified by the researcher. Researchers often evaluate multiple models with different numbers of topics, with a lack of established criteria for final selection. Numerous metrics for evaluating models quantitatively exist, although some researchers argue these are decoupled from subjective evaluations (Chang et al., 2009; DiMaggio, 2015; Lau et al., 2014; Roberts et al., 2016). In other words, models that look good quantitatively may not be ideal subjectively. The second challenge is labeling the topics once a model is selected. The label applied to a topic can impact how people make sense of the topic and its relationship to other topics or other variables. However, researchers may label the same topic in quite different ways, presenting a potential source of bias.

These challenges are often intertwined. For example, whether a topic makes sense subjectively is related to the difficulty in creating an appropriate label (Lee and Martin, 2015). Researchers often undertake these processes on their own. Other methods for evaluating the quality of topics and labels presuppose the researcher has already selected the number of topics and trained the model (Chang et al., 2009). Researchers may solicit opinions from domain experts or conduct studies to get opinions on topic labels for a model they have trained, but money and time present barriers to repeating this process for numerous models. Below we describe our approach to model selection and topic labeling.

Our first step in selecting the number of topics was to narrow the range of models we considered. Although it can be problematic to simply optimize an objective

function, it can be costly to manually evaluate every model, and topics may not be stable if training is repeated (Roberts et al., 2016). Beginning with a wide range of model sizes, we followed prior research (Roberts et al., 2014; Munoz-Najar Galvez et al., 2020) in using multiple automated means of evaluation as a first pass to reduce the space of possible models.

We considered three metrics to narrow this range: coherence, exclusivity, and perplexity. Semantic coherence refers to the frequency with which the top words in a topic appear together (Mimno et al., 2011). Exclusivity refers to how distinctive the top words are in each topic (Bischof and Airoldi, 2012). To the extent that the researcher wishes to use these metrics, it is desirable to maximize both coherence, which tends to occur with fewer topics, and exclusivity, which is maximized with more topics. There is thus a tradeoff between the two. The third metric we consider is perplexity, a measure used to assess how well the model fits held-out data. Lower perplexity suggests a better fit between the model and the underlying data but does not necessarily suggest more interpretable topics (Chang et al., 2009).

To identify a range of models to manually evaluate, we began by training models with three to 100 topics using the doParallel (Daniel et al., 2022), tm (Feinerer et al., 2008), and topicmodels (Grün and Hornik, 2011) libraries for R with Gibbs sampling. For each value of k, we evaluated perplexity using five-fold cross-validation while computing the average coherence and exclusivity of the training folds using the topicdoc library (Friedman, 2022). To assess the tradeoffs between these metrics, we standardized them so that the means are equal to zero and the standard deviations are equal to one. We identified a narrower range of values of k based on tradeoffs among these three metrics and trained models at three points in this range using the full corpus. Two authors not involved in the previous qualitative analysis then subjectively evaluated these three models based on the words most strongly associated with each topic. We preferred a parsimonious model that would be easy to describe. If we were to select a larger model, the topics would need to be sufficiently more informative to justify complicating subsequent analysis. To assess how robust our results are to model size, we selected an alternative model using another k in this range.

Labeling the topics

After training our preferred model, three authors independently labeled each topic based on the top 10 words and the three most relevant documents. One of these authors (MCH) was the lead author of the qualitative study. The two authors not involved in the prior study (ASM and SAS) then discussed these three sets of labels and assigned one label to each topic. These three authors then discussed these labels and made final changes.

Evaluating the topic model

We directly evaluate the extent to which our topics reflect the prior qualitative analysis (Halley et al., 2018) using a novel, simulation-based approach. Our simulation-based approach and comparison to codes from a qualitative analysis contrast with prior efforts to evaluate topic models, which often rely on discussion with domain experts, recruitment of human subjects, or automated means like coherence and exclusivity. To our knowledge, this formal approach is a novel contribution to the computational social science literature.

In the qualitative analysis (Halley et al., 2018), a given code (e.g. "Pay/Compensation") was either applied to a document or not. We can treat the application of a code to a document as a binary variable (0 = not applied, 1 = applied). In contrast, there is a continuous relationship between a document and a topic (how much the document is "about" the topic). To capture the strength of the relationship between these measures, we regressed each code on each topic using logistic regression as implemented in the scikit-learn library (Pedregosa et al., 2011). We then used the NumPy library (Harris et al., 2020) to compute the coefficient of discrimination (Tjur R^2) to measure the extent to which each topic explains the presence of each code (Tjur, 2009). This measure is equivalent to the more familiar coefficient of determination (R^2), but it can be calculated for logistic regression models. We then regressed each code on the complete set of topics using logistic regression to assess how well the full topic model explains variation in individual codes.

Below, we interpret the extent to which the topic model learns the same themes as the qualitative analysis, considering what the two approaches have in common and any themes that are specific to one approach.

Specifying an alternative model for contrast

We assessed the dependence of our results on the number of topics we selected by comparing our preferred and alternative models. Because the topics themselves are probability distributions over words, we assessed the relationships between the two sets of topics using Hellinger distances with the Gensim library (Řehůřek and Sojka, 2010). This measures the similarity between two probability distributions and ranges from 0 to 1. We calculated a measure of proximity using 1 minus the Hellinger distance.

We also compared our preferred and alternative models based on the prevalence of topics at the document level. In this comparison, documents are represented as probability distributions over topics. Because these comparisons focus on the probabilities of topics within these document vectors, rather than on probability distributions, the Hellinger distance is inappropriate. Instead, we use the Spearman rank-order correlation coefficient as implemented in the SciPy library for Python

(Virtanen et al., 2020). This is a nonparametric alternative to the more common Pearson correlation coefficient.

Simulations

To test whether the associations were stronger than we would expect by chance, we created one thousand synthetic corpora for comparison. For each corpus, we sampled document lengths (word counts) for 988 synthetic documents (being the number of documents in the real corpus) from the empirical distribution of document lengths. For each document, we sampled words from the empirical distribution of word frequencies. We then linked the synthetic documents to the qualitative codes, which we kept fixed so that the *i*th synthetic document in each synthetic corpus was associated with the same codes as the *i*th document in the real corpus, although the document length and content differed. We then trained two LDA models for each synthetic corpus using the numbers of topics in the preferred and alternative models. All other hyperparameters were the same as those used in the models trained on the real data. We then calculated Tjur R^2s by regressing each code on each topic and regressing each code on the full model for each simulated topic model.

Comparing real and simulated estimates

Our null hypothesis is that the amount of variance in a code that is explained by the prevalence of a single topic or by a full topic model via logistic regression (using the Tjur R^2) is random, conditional on the empirical distribution of document lengths and word frequencies. Having calculated Tjur R^2s for each simulated corpus, we compared each real estimate to the distribution of simulated estimates as in the bootstrap percentile method. Given the substantial number of estimates being compared, we corrected for multiple comparisons using the Bonferroni method.

Topic order is not fixed, so we did not require that the Tjur R^2 from regressing code_i on topic_j could only be compared to the Tjur R^2 from regressing code_i on topic_j in each simulated model. Instead, we allowed that the Tjur R^2 from regressing code_i on topic_j in one of the real models (preferred or alternative) could be compared to the Tjur R^2s from regressing code_i on any topic in each simulated topic model with the same number of topics. Our comparisons for the bivariate estimates are therefore to distributions of $k \times 1000$ simulated estimates. The Tjur R^2 from regressing code_i on a full topic model (preferred or alternative) can only be compared to one thousand estimates—one from each simulated model with the same number of topics. Thus, despite correcting our thresholds for significance using the Bonferroni method, we have enough simulated estimates to compare each real Tjur R^2 from a bivariate logistic regression to a distribution of simulated estimates using the percentile method for accepted thresholds of statistical significance. In contrast, applying the percentile method to each Tjur R^2 from regressing a code on one of a full topic model amounts to testing whether the Tjur R^2 is higher than all estimates from the simulated models with the same number of topics.

In our comparison of the topics in preferred and alternative models treating topics as probability distributions of words in the vocabulary, we compare each real estimate (1 minus the Hellinger distance) for topic_i in the preferred model and topic_j in the alternative model to the 1000 estimates based on the simulations.

Finally, in our comparison of the preferred and alternative models using the prevalence of topics at the document level, we compare the real Spearman rank-order correlation between the document probabilities of topic_i in the preferred model and topic_j in the alternative model to the 1000 corresponding coefficients from the models trained on the synthetic corpora.

Results

Model performance

Figure 1 depicts the trends in semantic coherence, exclusivity, and perplexity for models with values of k from 3 to 100. The measurements are standardized so they are on the same scale and were plotted using LOESS. This figure shows the tradeoff between coherence (the solid green line) and exclusivity (the dashed orange line). Until approximately $k = 35$, models show above-average coherence and below-average exclusivity. Near $k = 35$, models begin to favor exclusivity. Perplexity (the dashed purple line) has a quadratic relationship with k. In this modeling space, perplexity is minimized with fewer topics. Two authors (ASM and SAS) identified the region between $k = 15$ and $k = 35$ as the most promising based on these tradeoffs.

To probe these trends further, we assessed each metric's change from one value of k to the next using Figure 2. Rates of change in each metric seem to stabilize as the number of topics approaches 50.

Research Question 1: What types of discrimination are reported by physician mothers?

Table 2 presents the topics in our preferred model. For each topic, this table presents the final label, the 10 words most strongly associated with the topic, and a representative excerpt from a document strongly associated with the topic. There was immediate agreement among the authors about labeling some topics (e.g. "Family leave"), while other topics captured multiple themes. For example, documents closely related to Topic 15 (labeled "Pumping") also discussed issues such as disciplinary actions and

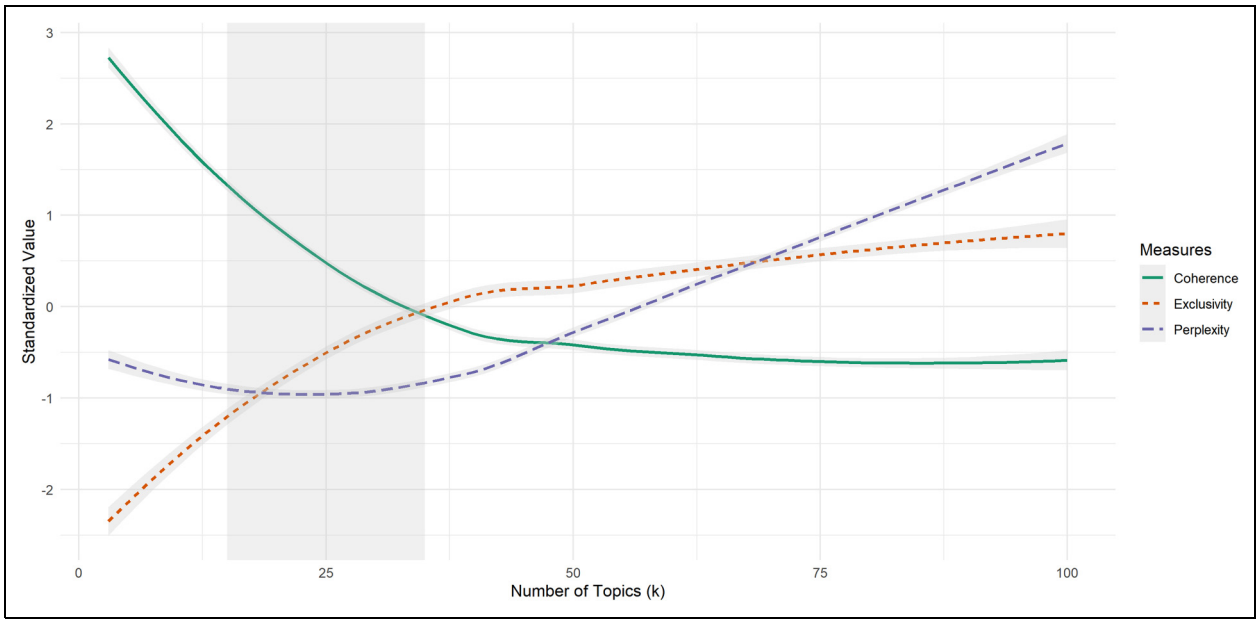

Figure 1. Tradeoffs among coherence, exclusivity, and perplexity over number of topics.

complaints from patients about physician mothers' need to pump. We chose the broader label "Pumping" for this topic because it more clearly distinguished the topic.

Research Question 2: Does topic modeling with LDA surface the same codes as human-generated thematic analysis?

Figure 3 illustrates the results of our quantitative comparison of the codes applied during the human-generated analysis and the topics from our preferred model ($k = 15$) at the document level using the coefficient of discrimination (Tjur, 2009). Because the Tjur R^2 from regressing code_i on topic_j in the preferred model could be compared to the Tjur R^2 from regressing code_i on any topic in the simulated models, each estimate from the bivariate regressions involving the preferred model was compared to a distribution of 15,000 simulated estimates (15 topics × 1000 simulations). Formal hypothesis testing was conducted using the bootstrap percentile method with the Bonferroni correction applied to the threshold for statistical significance. The Tjur R^2 for regressing code_i on the full model could only be compared to 1000 simulated estimates (i.e. one for each of the 1000 simulated models with the same number of topics), so using the percentile method with a Bonferroni-corrected significance threshold resulted in testing whether each Tjur R^2 for the full model was higher than each simulated estimate. To facilitate interpretation, we have removed non-significant coefficients and shaded significant cells to indicate how much variance in each code is explained by each topic or by the entire model.

In the bivariate analyses, the strongest associations are between the "Family Leave" code and the "Family leave" topic; the "Medical Training" code and the "Training" topic; the "Pay/Compensation" code and the "Career advancement" topic; and the "Childcare/Household Challenges" code and the "Burnout" topic. Interestingly, we find that the "Hierarchy" topic is associated with codes relating to both the workplace ("Job Changes") and the home ("Childcare/Household challenges").

On the other hand, we see no evidence of several expected associations. Although the "Job Changes" code is significantly associated with several topics ("Power," "Burnout," "Training," and "Staff interactions"), we find no evidence that this code is associated with the "Transitions," "Promotion inequality," or "Career advancement" topics. We also note that the topics to which we assigned more abstract labels are associated with relatively few codes: "Power" is only associated with the "Missed Opportunities" and "Job Changes" codes; "Agency" is only associated with the "Motherhood Specific" code; and "Respect" is not significantly associated with any codes.

Additionally, we do not find evidence that the "Expectations" or "Incentive/Pay Structure" codes are explained by any of our topics or by the entire preferred model better than chance. We also find that our "Transitions" topic is associated only with the "Academic Medicine" code, while the "Unequal compensation" topic—like "Agency"—is only associated with the "Motherhood Specific" code.

Notably, neither the individual topics nor the preferred model overall explains variance in the "Great quote/example" code better than chance. This may suggest the qualitative researchers were not biased toward particular

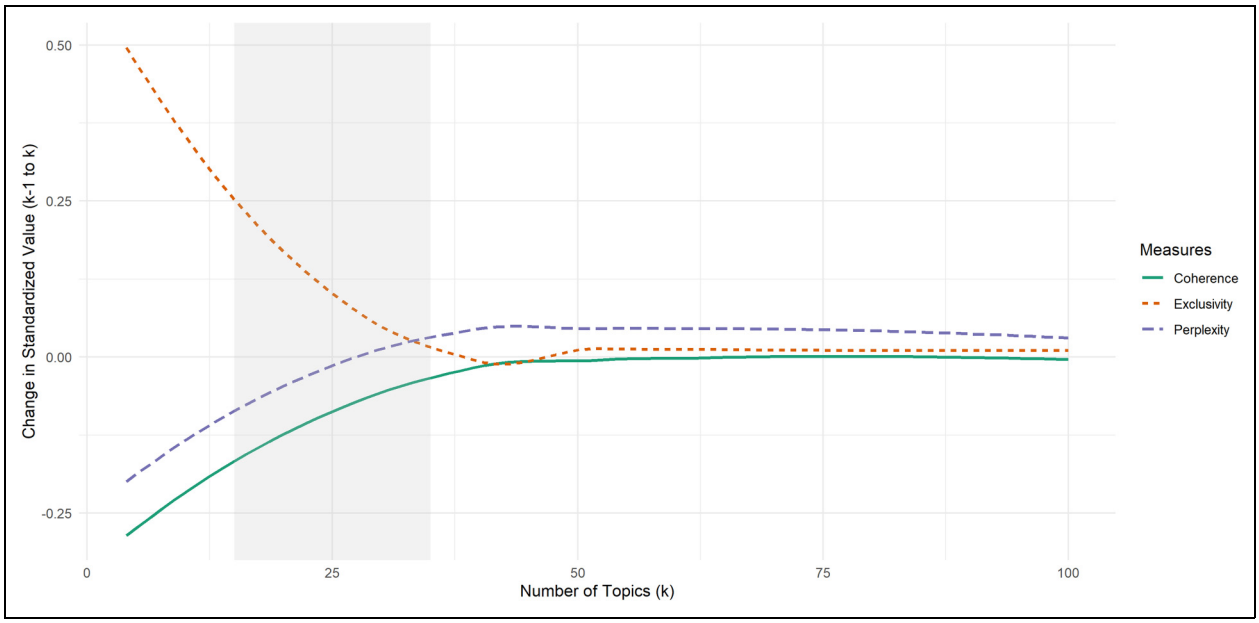

Figure 2. Change in coherence, exclusivity, and perplexity from $k-1$ to k Topics.

latent themes in their designation of what counted as a great quote or example.

Overall, our preferred model seems to surface most codes identified in the qualitative thematic analysis. Despite this, we do not see strong evidence that the topics generally explain much variance in the documents to which the codes were applied. As in the example above, the "Promotion inequality" topic explains little variance in codes such as "Missed Opportunities," although this code is associated with the "Power" and "Career advancement" topics.

Research Question 3: Is the comparability of computer-generated topics to human-generated codes robust to the number of topics the researcher chooses?

Three comparisons answer our third research question: a comparison of the topics in our preferred model ($k=15$) to the topics in the alternative model ($k=35$), a comparison of these models based on topic prevalence at the document level, and, finally, a comparison of the alternative model ($k=35$) to the qualitative coding. The robustness of our analysis to the number of topics we selected would be evident through either of two possibilities:

Possibility 1: The alternative model identifies the same 15 topics and 20 unrelated topics.

Possibility 2: The alternative model fragments the original 15 topics into distinct sets of more specific topics.

According to Possibility 1, the alternative model effectively encompasses the preferred model, providing the same 15 topics and then, because we specified that the model should have 35 topics, 20 additional topics. This possibility would be supported if two conditions obtain: first, each topic in the preferred model should be associated strongly with exactly one topic in the alternative model. Second, there should be 20 topics that are unrelated to any topic in the preferred model.

According to Possibility 2, on the other hand, each of the 15 topics in the preferred model should be strongly associated with at least one topic in the alternative model. The distinguishing feature of this hypothesis is that each topic in the preferred model may be associated with multiple topics in the alternative model; however, if one topic in the preferred model is associated with a given set of topics in the alternative model, no other topic in the preferred model should be associated strongly with any topic in that set.

Our first comparison treats each topic as a probability distribution over the words in the vocabulary. Each cell in Figure 4 represents the proximity of topic_i in the preferred model to topic_j in the alternative model using 1 minus the Hellinger distance. The estimate in each cell_{ij} was compared to the distribution of proximities in the corresponding cells from the 1000 simulations using the percentile method. Due to correcting for multiple comparisons, the result is that we test whether the real estimate is larger than all simulated estimates. Non-significant estimates are masked to improve interpretability. Blue cells represent distributions significantly closer than chance, and they are shaded according to proximity.

The relative strengths of the associations we observe could lend themselves to either possibility. Consistent with Possibility 1, seven of the 15 topics in the preferred model (rows) appear to be significantly associated with only one topic in the alternative model (columns). Further, 22 topics in the alternative model (62.9%) are

Table 2. Topic labels, representative words, and representative quotes.

Topic	Representative Words and Quotes
Topic 1	year fellowship research end support second leave change want institution
Transitions	"I had a very messy transition away from my old job at the university…"
Topic 2	maternity_leave leave week month pregnancy year find vacation contract academic
Family leave	"Residency/fellowship allowed just six weeks of maternity leave…"
Topic 3	woman man hire boss colleague meeting want young faculty bring
Promotion inequality	"My first chair and an academic institution gave the man I hired in the lower academic rank who I supervised the title of director of clinical services."
Topic 4	group practice partner hospital new clinic pay large board administrative
Power	"[M]et with a great deal of push-back likely because it means less cash for the male partners who would have to vote to share the wealth."
Topic 5	work hour kid work_time clinical home support family find husband
Burnout	"Unfortunately the productivity demands became so much that I could no longer focus an adequate amount of time on teaching and other non-clinical activities. We were incentivized only by our productivity in clinical settings"
Topic 6	job feel pay gender finally opportunity staff need late bonus
Unequal compensation	"A strong culture of silence around discussing pay. Female docs were expected to work harder and take on the tasks no-one really wanted."
Topic 7	like doctor good come help think want look know day
Psychosocial support	"I find there is absolutely no room for me in my life. I come behind everyone else even drop in patients I never agreed to care for."
Topic 8	physician thing decision experience need situation place way run think
Respect	"Bullying by nursing staff I feel is a norm in city hospitals in [region]…"
Topic 9	residency resident pregnant comment attending attend program chief surgery office
Training	"As a resident pumping for my 3 month old child […] I was told by my associate program director that my 'personal life was interfering with my ability to do [sic] perform my work responsibilities…'"
Topic 10	nurse female male_physician question female_physician age respect treat nursing refuse
Staff interactions	"One of the day shift charge nurses will routinely perform administrivia tasks for my male colleagues but refuses to perform them for me."
Topic 11	salary male position offer high promotion chair department training office
Career advancement	"Only when I saw that he was up for promotion, did I say 'hey shouldn't I be up for promotion.'"
Topic 12	discrimination peer base know old pay medical interview change care
Favoritism	"Good old boy hospitalists and surgeons didn't follow my suggestions…"
Topic 13	tell ask male_colleague director hold leadership male way admin administration
Hierarchy	"My director did not even consider me for the position of assist director."
Topic 14	time child schedule care allow decision day shift actually benefit
Agency	"Generally my sense of burnout and discrimination revolves around lack of autonomy and say in how my day is structured and policies around patient care."
Topic 15	patient staff pump doctor room start issue administration complaint support_staff
Pumping	"In residency I was constantly targeted for having to pump milk."

not significantly associated with any topic in the preferred model. This is close to the 20 unrelated themes predicted by Possibility 1. However, one in three topics in the preferred model ("Agency," "Favoritism," "Respect," "Unequal compensation," and "Promotion inequality") is not associated with any topic in the alternative model.

We also see patterns consistent with Possibility 2. The "Pumping," "Career advancement," and "Unequal compensation" topics are each associated with two topics in the alternative, and these topics, in turn, are not associated with any other topic in the preferred model. This result is also consistent with our observation that the topic we labeled "Pumping" seemed to have multiple facets. The larger model may have fragmented the "Pumping" topic. In our comparison of the top words and most relevant documents for each topic, "Pumping" and Topic 25 seem similar. However, Topic 25 does not appear to reflect some aspects of the "Pumping" topic in the preferred model, like disciplinary actions and patient complaints, so it does appear to be more specific.

Comparing the preferred and alternative models at the document level

While the previous comparison focuses on the proximity of topics as probability distributions over words, our second

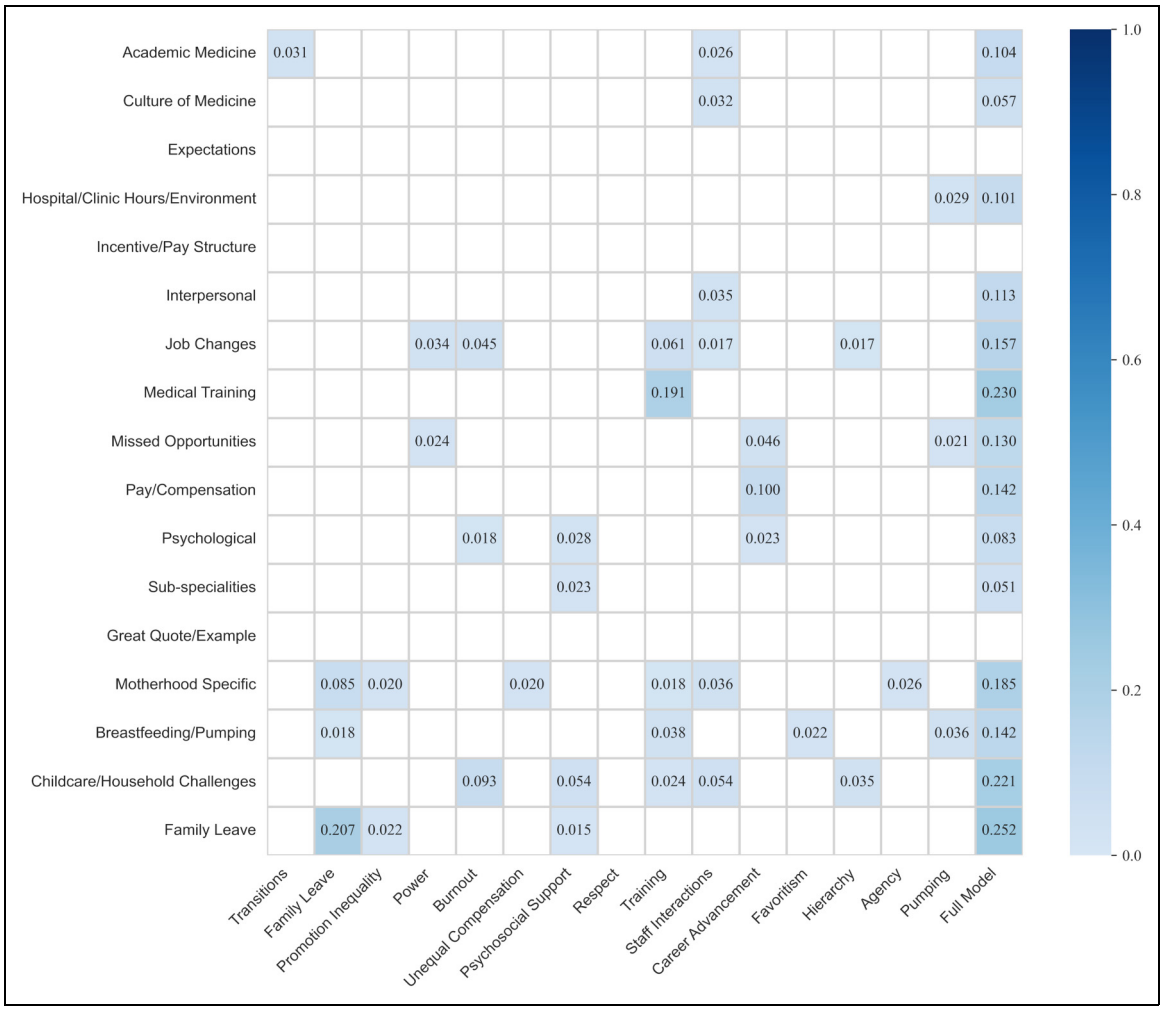

Figure 3. Coefficients of discrimination between codes and topics in preferred model ($k = 15$).

comparison of our preferred and alternative models focuses on the document-level prevalence of topics. We compare the document probabilities of topic_i from the preferred model to the document probabilities of topic_j using the Spearman rank-order correlation coefficient and compare the real coefficients to the corresponding 1000 coefficients from the simulated models. Using the bootstrap percentile method with a Bonferroni-corrected significance threshold, this approach tests whether the real correlation coefficients are more extreme than all coefficients from the simulations.

Figure 5 presents the results of our second comparison with non-significant coefficients masked to aid interpretation. This document-level analysis largely replicates the preceding analysis, finding 10 of 13 relationships (76.9%) but missing the associations between "Psychosocial support" and Topic 21, "Pumping" and Topic 25, and "Hierarchy" and Topic 35. "Pumping" is again associated with Topic 30. However, this comparison found nine additional significant associations, filling in some of the gaps previously identified: significant relationships were found between "Favoritism" and Topic 35, "Respect" and Topic 15, "Promotion inequality" and Topic 11, and "Unequal compensation" and each of Topics 20, 22, and 32. Only two topics in the preferred model are unrelated to topics in the alternative model according to this analysis, namely "Agency" and "Psychosocial support." "Power" is most strongly associated with Topic 8, as in Figure 4, but is also related to Topic 17. Similarly, "Staff interactions" remains most strongly associated with Topic 2 but is also related to Topic 3. "Career advancement" was found to be negatively correlated with Topic 30 in this analysis.

Comparing the alternative model to the qualitative analysis

Figure 6 presents the Tjur R^2 from regressing each code from the qualitative analysis on each topic individually (columns 1–35) and then on all 35 topics in the alternative model (column 36). Each bivariate estimate is compared to

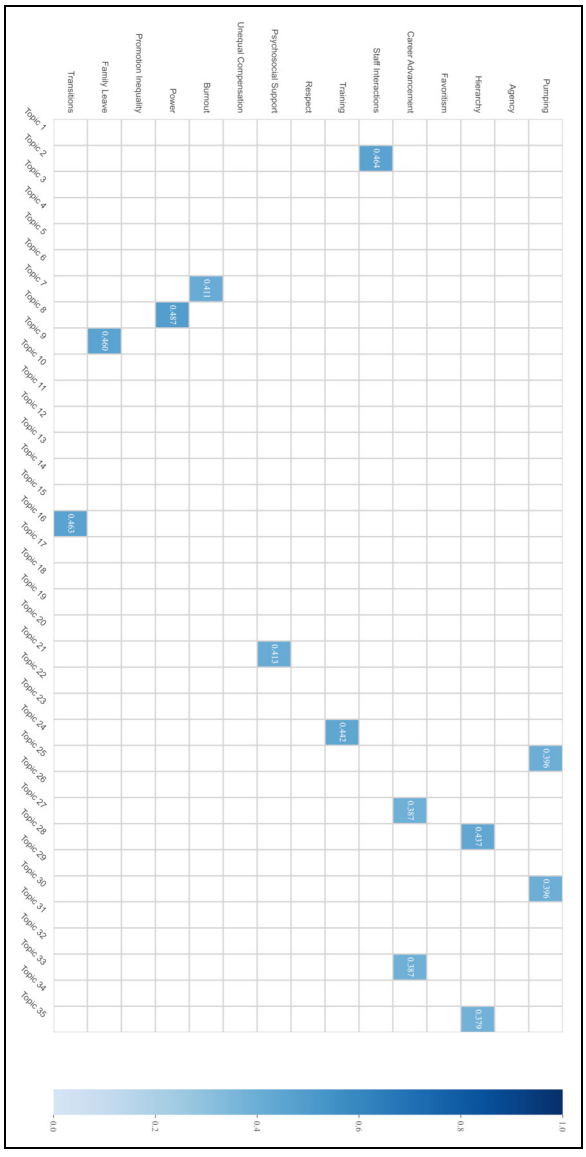

Figure 4. Proximity (1 – Hellinger Distance) of topics as distributions over words in the preferred ($k = 15$) and alternative ($k = 35$) models.

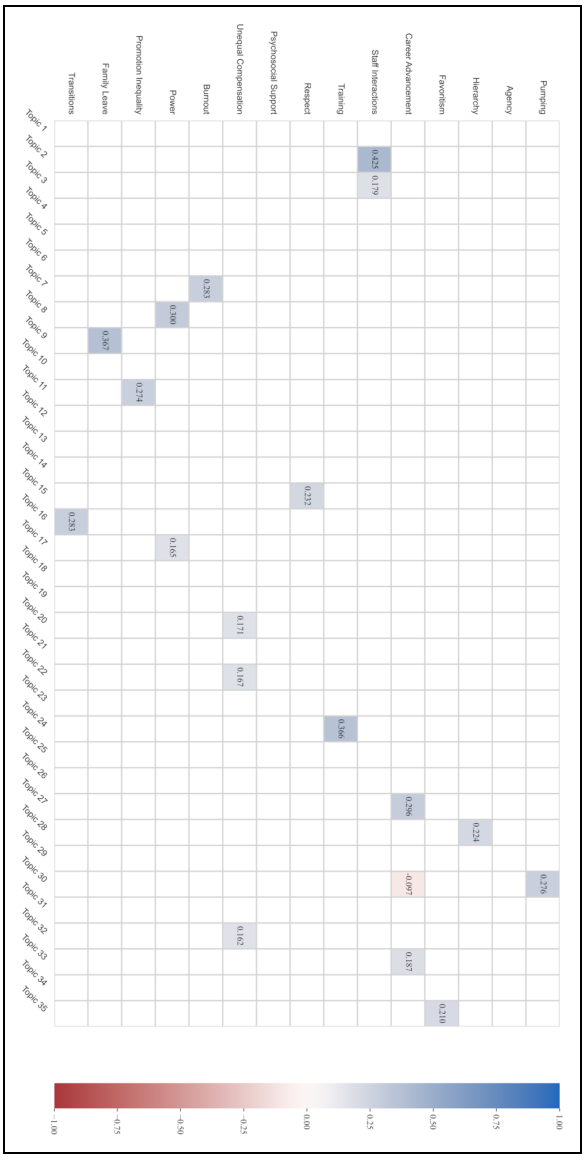

Figure 5. Spearman rank-order correlation coefficients between topic probabilities at the document level in the preferred ($k = 15$) and alternative ($k = 35$) models.

a distribution of 35,000 estimates (35 topics × 1000 simulations), while each estimate from regressing a code on the full alternative model is compared to the 1000 corresponding simulated estimates. As in Figures 3 and 4, blue cells indicate significant differences based on comparisons to the distribution of simulated estimates using the percentile method with the Bonferroni correction for multiple comparisons.

Most topics (21 topics, or 60%) in the alternative model are not associated with any of the qualitative codes. Further, the alternative model fails to explain variance in the "Culture of Medicine," "Expectations," and "Incentive/Pay Structure" codes. In contrast, the preferred model ($k = 15$) did explain variance in the "Culture of Medicine" code.

A few other findings point to consistency among these approaches. First, the "Family Leave" code is only associated with Topic 9 in the alternative model (Figure 6), and Topic 9 is also only associated with the "Family leave" topic in the preferred model (Figures 4 and 5). The "Medical Training" code is only associated with Topic 24 in the alternative model (Figure 6), which is, in turn, only associated with the "Training" topic in the preferred model (Figures 4 and 5). The "Breastfeeding/Pumping" code is associated with Topics 9, 25, and 27 in the alternative model (Figure 6). The "Pumping" topic in the preferred model is similarly associated with Topic 25 (Figures 4 and 5)—but also with Topic 30 (Figure 4). Whereas the "Breastfeeding/Pumping" code is weakly associated with four topics in the preferred model

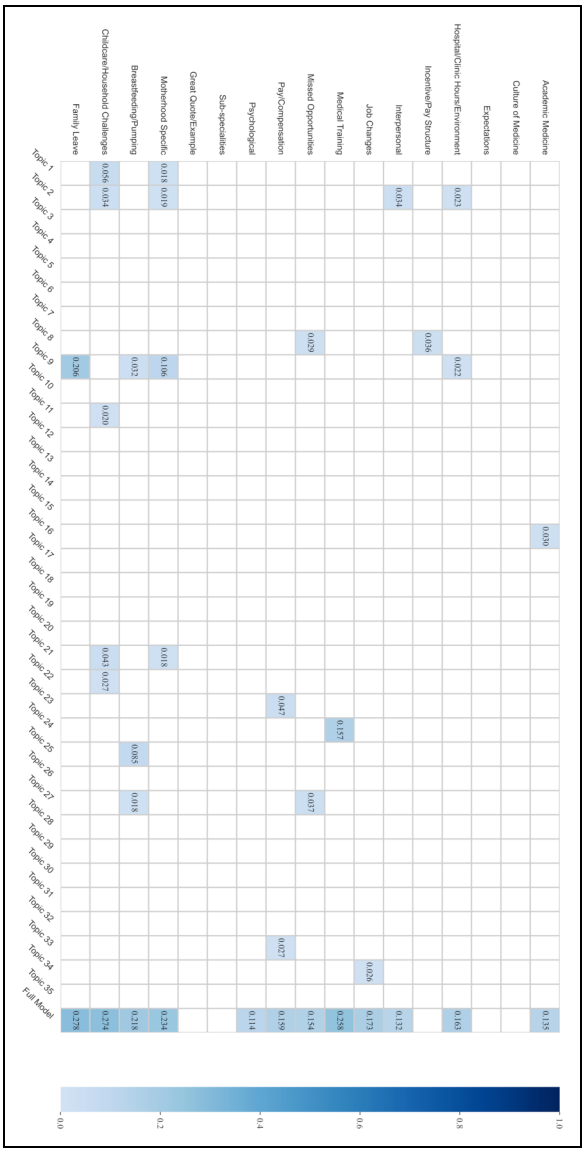

Figure 6. Coefficients of discrimination between codes and topics in alternative model ($k = 35$).

("Family leave," "Training," "Favoritism," and "Pumping") but not well-explained by the full preferred model (Tjur $R^2 = 0.142$), it is more strongly associated with Topic 25 and is better explained by the full alternative model (Tjur $R^2 = 0.218$).

Additionally, Topic 2 in the alternative model is associated with the "Hospital/Clinic Hours and Environment," "Interpersonal," "Motherhood Specific," and "Childcare/Household Challenges" codes (Figure 6). Topic 2 is also associated with the "Staff interactions" topic in the preferred model (Figures 4 and 5), which is similarly associated with the "Interpersonal," "Motherhood Specific," and "Childcare/Household Challenges" codes (Figure 3).

Table 3. Explanation of variance in qualitative codes by preferred ($k = 15$) and alternative ($k = 35$) LDA topic models.

Code	Preferred Model (Tjur R^2)	Alternative Model (Tjur R^2)	R^2 Ratio
Academic Medicine	0.104*	0.135*	0.77
Culture of Medicine	0.057*	0.066	0.86
Expectations	0.042	0.056	0.75
Hospital/Clinic Hours and Environment	0.101*	0.163*	0.62
Incentive/Pay Structure	0.038	0.098	0.39
Interpersonal	0.113*	0.132*	0.86
Job Changes	0.157*	0.173*	0.91
Medical Training	0.230*	0.258*	0.89
Missed Opportunities	0.130*	0.154*	0.84
Pay/Compensation	0.142*	0.159*	0.89
Psychological	0.083*	0.114*	0.73
Sub-specialities	0.051*	0.069	0.74
Great Quote/Example	0.031	0.105	0.30
Motherhood Specific	0.185*	0.234*	0.79
Breastfeeding/Pumping	0.142*	0.218*	0.65
Childcare/Household Challenges	0.221*	0.274*	0.81
Family Leave	0.252*	0.278*	0.91

Comparing the full models

Table 3 presents the coefficients of discrimination (Tjur R^2) for the models regressing each code on the full preferred and alternative models. The final column in Table 3 presents the ratio of the Tjur R^2 from the preferred model to that of the alternative model. As expected, the larger model explains more variance in each code. However, a few things stand out. First, despite the larger Tjur R^2 (R^2 ratio = 0.39), the alternative model does not seem to explain variance in the "Incentive/Pay Structure" code better than chance. Second, although the larger model nominally explains more variance in the "Culture of Medicine" code (Tjur $R^2 = 0.066$), it does not differ from chance; the preferred model does explain this code significantly better than chance (Tjur $R^2 = 0.057$). We see the same pattern for the "Sub-specialities" code, which is significantly associated with the "Psychosocial support" topic in the preferred model (Tjur $R^2 = 0.023$; full model Tjur $R^2 = 0.051$) but is not explained better than chance by the larger model (Tjur $R^2 = 0.069$). Otherwise, the models exhibit similar patterns of association with the qualitative codes.

Discussion

Despite the growth of computational social science, the relationship between machine- and human-generated thematic text analysis remains poorly described, and tradeoffs between methods remain vague. In this study, we bridge

qualitative and computational methods through two steps. First, we formally compare topic modeling and qualitative thematic analysis using a corpus of physician mothers' experiences of workplace discrimination (Research Questions 1 and 2). Second, we demonstrate the effects of modeling decisions on translation between computationally driven and human-driven themes (Research Question 3).

Comparing computational to human-generated themes

The computationally derived topics from the preferred model seem to capture many of the themes identified in the qualitative analysis. The topic models and qualitative analysis triangulate specific themes, most notably family leave, medical training, and breastfeeding/pumping. Each topic model has a topic that is simultaneously associated with the "Interpersonal," "Motherhood Specific," and "Childcare/Household Challenges" codes ("Staff interactions" and Topic 2 in the preferred and alternative models, respectively). This suggests that the models detect similar patterns of association among themes.

The topic models explain variance in several codes from the qualitative analysis quite well. Interestingly, the different approaches may have identified similar themes based on different documents. This points to the usefulness of inductively deriving themes that can then be analyzed using other methods. From this, we might conclude that our approach to topic modeling is useful for identifying potential codes for qualitative analyses, if not necessarily for *applying* them.

In general, the degree of correspondence between our topics and the qualitative coding should encourage researchers who are unsure about investing resources in collecting such data as part of surveys or other online studies (Roberts et al., 2014).

Comparisons across model sizes

We find our comparisons of preferred and alternative models promising. Although LDA and similar approaches are sensitive to modeling decisions, both topic models picked up many of the same broad themes that human coders surfaced. This suggests topic modeling may be a useful first appraisal of potential themes in a dataset. A first look like this may be useful for an iterative approach to labeling or subsetting data in a qualitative context or for labeling data for a supervised learning problem. This may make analyses feasible for larger datasets, such as those gathered from regional or global social media-based surveys.

Our comparisons of the preferred and alternative models show that the larger model does not merely identify the smaller model's topics and 20 additional topics. Although we do see that specific topics in the larger model seem to correspond well to single topics in the smaller model (consistent with Possibility 1), we also see that the larger model, in some cases, seems to split our original topics into multiple topics (consistent with Possibility 2). This suggests that the number of topics is impactful and lower numbers do not merely eliminate less important topics. Selecting a larger k may result in splitting overly general topics into more distinct elements or dealing better with tokens that could have divergent interpretations (e.g. homonyms, metonyms, or polysemous words).

Methodological insights

Preprocessing. Although sometimes glossed over, the effects of preprocessing decisions are difficult to envision for those without computational experience. Table 1 illustrates that computational approaches involve "reading" text differently than a human would. Clarifying what an algorithm would "see" in contrast to what a human would interpret is a crucial step toward transparent and interpretable models and increasing readers' trust in computationally driven findings.

Computational applications. Extrinsic and intrinsic evaluation of the topic quality and overall model quality is a critical area of research, yet it is quite a different endeavor to evaluate topics according to whether humans would have identified them through a rigorous coding process. Our formal comparison of topic prevalence based on unsupervised topic models to the application of codes by trained human coders is a novel test of the quality of topics and topic models. Researchers using computational methods should find our results encouraging: our middle-of-the-road application of a widely used approach to topic modeling did a fair job of recovering many of the topics that were salient to a team of qualitative researchers in addition to foregrounding themes like power, hierarchy, favoritism, and respect.

Qualitative applications. Our findings provide insights for qualitative researchers looking to diversify their toolsets. Given that there are many different approaches to text analysis, even within the large "qualitative methods" umbrella, integrating computational methods may be more appropriate for certain approaches. More concrete codes (e.g. "Family Leave" and "Medical Training") seem more likely to correlate with topics generated computationally. This may suggest that computational approaches may be more useful when one is looking to use topic modeling to develop codes that directly reflect the participant's language (i.e. an "emic" interpretation), as opposed to identifying more interpretive, "etic" constructs that rely on examining not only the language itself but also how the language is used to make meaning out of individual or social experiences.

Further, computational approaches may be suited for identifying taken-for-granted or missed themes like the "Respect" topic, which had no overt counterpart in the qualitative codebook. Additionally, having a computer quantify obvious but unarticulated topics could be useful for downstream tasks or as verification of code framing. While this comparison shows promise for the intersection of human and computational approaches to text analysis, further research is needed to examine how these methods intersect to better guide their integration in specific projects.

Although topic models and other computational methods may be sensitive to researchers' assumptions and model specification decisions, they have the benefit of identifying many similar themes in a much less costly way. The qualitative analysis to which we compare our models was a time-intensive process involving the labor of a team of experts; coordinating such efforts can be complex and may be infeasible with large datasets. The overlap between the qualitative analysis and the patterns identified in our topic models is encouraging. At the same time, topics from LDA are fundamentally probability distributions over words; whether we judge them as interpretively satisfying "themes" may vary. In a study of figurative language, Rhody (2012) discusses a typology of topics based on how patterns in different documents can influence LDA. Ylä-Anttila and colleagues (2022) provide insight into the conditions under which topics can be said to be like the frames of frame analysis.

Limitations. An obvious limitation of this study is that we evaluate only one approach to unsupervised computational text analysis. Different data, preprocessing, model specifications, or methods could produce results with different relationships to human coding of the same data. This merits further evaluation. The application of deep learning to topic modeling, for example, has resulted in an increasing variety of alternatives to established approaches like LDA (Zhao et al., 2021). Our corpus is also relatively small for topic modeling. Further, although our approach to evaluating and labeling topics is not uncommon, interactive tools like LDAvis (Sievert and Shirley, 2014) can improve model evaluation and interpretation. While our conclusions cannot speak to the relationship between computational and qualitative approaches *in general*, they suggest a bridge between methods that, when thoughtfully deployed, can address the limitations of either approach.

We also only compare our computational approach to one approach to qualitative coding. Explicating the relationships between different quantitative and qualitative approaches will require many iterations across different contexts. We also compare our topic models to the application of the qualitative codes at the document level. This comparison permits a rigorous formal analysis, but it excludes the higher-order themes and mechanisms explored by the qualitative team (Halley et al., 2018). It may be the case that the latent themes we identify computationally are not comparable to the higher-order themes that the qualitative researchers organized the codes into, as Baumer and colleagues (2017) also suggest.

Further, although we used measures of coherence, exclusivity, and perplexity at one stage of our model selection process, we do not speak to whether models with higher values of these metrics better reflect human identification of themes in text or whether different metrics (e.g. different versions of coherence) are preferable. We share a wariness of focusing on automated model evaluation with the researchers we cite above and note that we used tradeoffs among multiple approaches to narrow the range of candidate model sizes. This approach is in line with thoughtful work on these problems (Roberts et al., 2016).

Notably, there is no one-size-fits-all approach to preprocessing and text normalization. For example, a review by Hickman and colleagues (2022) provides a flow chart but discusses various complications. Researchers sometimes assume algorithms like LDA obviate the need for steps like lemmatization and identifying *n*-grams. In contrast, we feel that forgoing these steps places considerable faith in the algorithm. Identifying *n*-grams disambiguates word senses, for example, removing uncertainty about the meaning of "leave" in the domain-relevant bigram "family_leave." Without this step, a model may still learn that the words "family" and "leave" are related, but all occurrences of "family" and "leave" (even where they have other meanings) will be assumed to relate to the same topics in the same ways. Treating "family leave," "family," and "leave" as distinct types eliminates this issue and clarifies that "family_leave" has a policy-oriented meaning that "family" and "leave" lack on their own.

Researchers have also argued that stemming and lemmatization are more appropriate for morphologically rich languages (Boyd-Graber et al., 2014; May et al., 2019). Text in the same language may vary in morphological richness by domain. Given our domain (i.e. medicine as a profession), the small size of our corpus, and our assumptions about language, we feel our preprocessing steps are appropriate. Nonetheless, readers may question how much our results stem from LDA itself versus our analytical decisions. We argue that if these steps were unnecessary or even harmful to our models, the counterfactual models should be equivalent or better.

Future directions. Experiences of discrimination are not limited to one group of people at one point in time. The hope for computational methods is that they may shorten the lag between developing questions and answering them. It is crucial that we better understand the dynamics of discrimination in the workplace, and this will not be accomplished by a single study (Deardorff and Dahl, 2016; Ridgeway, 2011). Directions to extend in are evaluations of fathers' experiences, same-sex couples, and

intersectional issues arising at the crossroads of gender, race, occupation, income, and age. Improved methods to understand experiences of workplace discrimination may allow institutions to design appropriate countermeasures and evaluate progress more quickly (Schiebinger, 2021).

The Internet and social media present novel avenues to gather experiences from marginalized groups. With access to thousands or millions of participants, traditionally human-driven document review is not feasible, motivating computationally driven knowledge creation. We suggest that topic modeling appears feasible for augmenting human-driven coding to understand workplace culture and discrimination. This approach answers calls for transparency about methods in computational social science (Jarrahi et al., 2021).

Summary. We have argued that improved methods for comparing quantitative and qualitative inductive coding will aid multiple stakeholders. This matters because concerns have been raised around reproducibility in social science (Tannenbaum et al., 2019; Wallach et al., 2018). In line with the current focus on transparency and model inspectability, here we make explicit our assumptions, test multiple approaches to model generation, and provide our code. Used with contextual awareness, topic models may provide a scalable and inspectable approach to supporting qualitative research methods. By establishing how well unsupervised approaches approximate what skilled humans do, we provide a useful starting point for future work identifying and refining latent themes in text data. These findings point to the merits of computational approaches for identifying meaningful themes and relationships in data collected in ways that should appeal to both quantitative and qualitative researchers.

Acknowledgements

The authors wish to extend their gratitude to Michelle Jackson, Daniel McFarland, Mark Hoffman, Klint Kanopka, Austin van Loon, and members of the Computational Sociology workshop at Stanford University for their feedback in response to various drafts and substantive questions pertaining to this project. They also wish to thank Anna K. M. Skarpelis, Marshall A. Taylor, and Matt Rafalow for their feedback and for the opportunity to present an earlier version of this paper at the annual meeting of the American Sociological Association.

Code and data availability

Our code is available at https://doi.org/10.5281/zenodo.7486093. Data are sensitive and may contain identifying information. Data may be available upon request subject to IRB oversight.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: ASM was supported by grants from the National Institutes of Health, National Center for Advancing Translational Science, Clinical and Translational Science Award (KL2TR001083 and UL1TR001085), and the Stanford Human-Centered AI Institute. MCH was supported by grants from the National Center for Advancing Translational Science, Clinical and Translational Science Award (UL1TR003142) and the National Human Genome Research Institute (K01HG011341). EL was supported by grants from the National Cancer Institute of the National Institutes of Health (DP2CA225433) and the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health (K24AR075060). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ORCID iDs

Adam S Miner https://orcid.org/0000-0002-5125-4735
Sheridan A Stewart https://orcid.org/0000-0001-5789-8390
Meghan C Halley https://orcid.org/0000-0002-5031-9840
Laura K Nelson https://orcid.org/0000-0001-8948-300X
Eleni Linos https://orcid.org/0000-0002-5856-6301

References

Adesoye T, Mangurian C, Choo EK, et al. (2017) Perceived discrimination experienced by physician mothers and desired workplace changes: A cross-sectional survey. *JAMA Internal Medicine* 177(7): 1033–1036.

Aisenbrey S, Evertsson M and Grunow D (2009) Is there a career penalty for mothers' time out? A comparison of Germany, Sweden and the United States. *Social Forces* 88(2): 573–605.

Arnold CW, Oh A, Chen S, et al. (2016) Evaluating topic model interpretability from a primary care physician perspective. *Computer Methods and Programs in Biomedicine* 124: 67–75.

Baumer EPS, Mimno D, Guha S, et al. (2017) Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal for the Association of Information Science & Technology* 68(6): 1397–1410.

Bird S, Loper E and Klein E (2009) *Natural Language Processing with Python*. Sebastopol: O'Reilly Media Inc.

Bischof J and Airoldi EM (2012) Summarizing topical content with word frequency and exclusivity. In: Proceedings of the 29th International Conference on Machine Learning, pp.201–208.

Blei DM (2012) Probabilistic topic models. *Communications of the ACM* 55(4): 77–84.

Blei DM and Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, pp.113–120.

Blei DM and Lafferty JD (2007) A correlated topic model of *Science*. *The Annals of Applied Statistics* 1(1): 17–35.

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Boyd-Graber J, Mimno D, Newman D (2014) Care and feeding of topic models: Problems, diagnostics, and improvements. In: Airoldi EM, Blei DM, Erosheva EA, et al (eds) *Handbook of*

Mixed Membership Models and Their Applications. New York: CRC Press, pp.225–254.

Chang J, Gerrish S, Wang C, et al. (2009) Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems* 22: 288–296.

Correll SJ, Benard S and Paik I (2007) Getting a job: Is there a motherhood penalty? *American Journal of Sociology* 112(5): 1297–1339.

Deardorff MD and Dahl JG (2016) *Pregnancy Discrimination and the American Worker*. New York: Palgrave Macmillan.

Denny MJ and Spirling A (2018) Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis* 26(2): 168–189.

DiMaggio P (2015) Adapting computational text analysis to social science (and vice versa). *Big Data & Society* 2(2): 1–5.

Feinerer I, Hornik K and Meyer D (2008) Text mining infrastructure in R. *Journal of Statistical Software* 25(5): 1–54.

Friedman D (2022) topicdoc: Topic-specific diagnostics for LDA and CTM topic models. https://cran.r-project.org/web/packages/topicdoc/index.html.

Gaut G, Steyvers M, Imel ZE, et al. (2017) Content coding of psychotherapy transcripts using labeled topic models. *IEEE Journal of Biomedical and Health Informatics* 21(2): 476–487.

Gerlach M, Peixoto TP and Altmann EG (2018) A network approach to topic models. *Science Advances* 4(7): eaaq1360.

Gerlach M, Shi H and Amaral LAN (2019) A universal information theoretic approach to the identification of stopwords. *Nature Machine Intelligence* 1(12): 606–612.

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

Grün B and Hornik K (2011) Topicmodels: An R package for fitting topic models. *Journal of Statistical Software* 40(13): 1–30.

Halley MC, Rustagi AS, Torres JS, et al. (2018) Physician mothers' experience of workplace discrimination: A qualitative analysis. *BMJ* 363: k4926.

Harris CR, Millman KJ, Van der Walt SJ, et al. (2020) Array programming with NumPy. *Nature* 585(7825): 357–362.

Hickman L, Thapa S, Tay L, et al. (2022) Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods* 25(1): 114–146.

Honnibal M, Montani I, Van Landeghem S, et al. (2020) spaCy: Industrial-strength natural language processing in Python. https://spacy.io/.

Hoyle A, Goel P, Hian-Cheong A, et al. (2021) Is automated topic model evaluation broken? The incoherence of coherence. *Advances in Neural Information Processing Systems* 34: 1–16.

Jarrahi MH, Newlands G, Lee MK, et al. (2021) Algorithmic management in a work context. *Big Data & Society* 8(2): 1–14.

Jena AB, Khullar D, Ho A, et al. (2015) Sex differences in academic rank in US medical schools in 2014. *JAMA* 314(11): 1149–1158.

Jena AB, Olenski AR and Blumenthal DM (2016) Sex differences in physician salary in US public medical schools. *JAMA Internal Medicine* 176(9): 1294–1304.

Krouska A, Troussas C and Virvou M (2016) The effect of preprocessing techniques on twitter sentiment analysis. In: 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA), pp.1–5.

Lau JH, Newman D and Baldwin T (2014) Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp.530–539.

Lee M and Martin JL (2015) Coding, counting and cultural cartography. *American Journal of Cultural Sociology* 3(1): 1–33.

Lehman LW, Saeed M, Long W, et al. (2012) Risk stratification of ICU patients using topic models inferred from unstructured progress notes. In: AMIA Annual Symposium Proceedings 2012, pp.505–511.

Lossio-Ventura JA, Lee AY, Hancock JT, et al. (2021) Identifying silver linings during the pandemic through natural language processing. *Frontiers in Psychology* 12: 712111.

May C, Cotterell R and Van Durme B (2019) An analysis of lemmatization on topic models of morphologically rich language. arXiv:1608.03995v2.

Melville S, Eccles K and Yasseri T (2019) Topic modeling of everyday sexism project entries. *Frontiers in Digital Humanities* 5(28): 1–9.

Daniel F, Microsoft Corporation, Weston S and Tenenbaum D (2022) doParallel: Foreach parallel adaptor for the 'parallel' package. https://cran.r-project.org/web/packages/doParallel/index.html.

Mimno D, Wallach HM, Talley E, et al. (2011) Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp.262–272.

Munoz-Najar Galvez S, Heiberger R and McFarland D (2020) Paradigm wars revisited: A cartography of graduate research in the field of education (1980–2010). *American Educational Research Journal* 57(2): 612–652.

Nelson LK (2020) Computational grounded theory: A methodological framework. *Sociological Methods & Research* 49(1): 3–42.

Nelson LK, Burk D, Knudsen M, et al. (2021) The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research* 50(1): 202–237.

Paul MJ and Dredze M (2014) Discovering health topics in social media using topic models. *PLOS ONE* 9(8): e103408.

Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(85): 2825–2830.

Ramage D, Hall D, Nallapati R, et al. (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp.248–256.

Ramage D, Manning CD and Dumais S (2011) Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.457–465.

Řehůřek R and Sojka P (2010) Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50.

Rhody LM (2012) Topic modeling and figurative language. *Journal of Digital Humanities* 2(1).

Ridgeway CL (2011) *Framed by Gender: How Gender Inequality Persists in the Modern World*. Oxford: Oxford University Press.

Roberts ME, Stewart BM and Tingley D (2016) Navigating the local modes of big data: The case of topic models. In: Alvarez RM (ed) *Computational Social Science: Discovery and Prediction*. New York: Cambridge University Press, pp.51–97.

Roberts ME, Stewart BM, Tingley D, et al. (2014) Structural topic models for open-ended survey responses. *American Journal of Political Science* 58(4): 1064–1082.

Röder M, Both A and Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp.399–408.

Sahlgren M (2008) The distributional hypothesis. *Italian Journal of Linguistics* 20: 33–53.

Schiebinger L (2021) Analyzing Research Priorities and Potential Outcomes, Gendered Innovations. Available at: https://genderedinnovations.stanford.edu/methods/priorities.html (accessed 5 October 2021).

Schofield A, Magnusson M, Thompson L, et al. (2017) Understanding text pre-processing for latent Dirichlet allocation. ACL Workshop for Women in NLP.

Schofield A and Mimno D (2016) Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics* 4: 287–300.

Schweinberger M, Haugh M and Hames S (2021) Analysing discourse around COVID-19 in the Australian Twittersphere: A real-time corpus-based analysis. *Big Data & Society* 8(1): 1–17.

Sievert C and Shirley KE (2014) LDAvis: A method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp.63–70.

Tannenbaum C, Ellis RP, Eyssel F, et al. (2019) Sex and gender analysis improves science and engineering. *Nature* 575: 137–146.

Tjur T (2009) Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *The American Statistician* 63(4): 366–372.

Virtanen P, Gommers R, Oliphant TE, et al. (2020) Scipy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* 17(3): 261–272.

Wallach JD, Boyack KW and Ioannidis JPA (2018) Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *PLOS Biology* 16(11): e2006930.

Ylä-Anttila T, Eranti V and Kukkonen A (2022) Topic modeling for frame analysis: A study of media debates on climate change in India and USA. *Global Media & Communication* 18(1): 91–112.

Yu WH and Kuo JCL (2017) The motherhood wage penalty by work conditions: How do occupational characteristics hinder or empower mothers? *American Sociological Review* 82(4): 744–769.

Zhao H, Phung D, Huynh V, et al. (2021) Topic modelling meets deep neural networks: A survey. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21) Survey Track, pp.4713–4720.