A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews

Byron C Wallace,1 Michael J Paul,2 Urmimala Sarkar,3 Thomas A Trikalinos,1 Mark Dredze4

ABSTRACT
Online physician reviews are a massive and potentially rich source of information capturing patient sentiment regarding healthcare. We analyze a corpus comprising nearly 60 000 such reviews with a state-of-the-art probabilistic model of text. We describe a probabilistic generative model that captures latent sentiment across aspects of care (eg, interpersonal manner). We target specific aspects by leveraging a small set of manually annotated reviews. We perform regression analysis to assess whether model output improves correlation with state-level measures of healthcare. We report both qualitative and quantitative results. Model output correlates with state-level measures of quality healthcare, including patient likelihood of visiting their primary care physician within 14 days of discharge (p=0.03), and using the proposed model better predicts this outcome (p=0.10). We find similar results for healthcare expenditure. Generative models of text can recover important information from online physician reviews, facilitating large-scale analyses of such reviews.

INTRODUCTION
Individuals are increasingly turning to the web for healthcare information. Indeed, a recent survey found that 72% of internet users have looked online for health information in the past year. One in five of these users have looked for reviews of either particular treatments or doctors. Although initial data revealed a paucity of doctor reviews online, a recent study of a random sample of 500 urologists found online reviews for about 80% of them.

People are not only consuming health information online: they are also producing it. This shift has generated a proliferation of health-related user-generated content, including online doctor reviews. Analyzing large corpora of such reviews may reveal interesting trends in consumer sentiment regarding their healthcare experiences.

Qualitative analyses of reviews can provide important insights, but require trained investigators to read and analyze text, and thus tend to be modest in size. Quantitative approaches can leverage the massive volume of textual data on the internet. Such methods may allow us to ‘harness the cloud of patient experience’ online. But they must be designed to capture the desired latent structure.

To this end, we utilize a state-of-the-art probabilistic model that jointly captures latent aspects and sentiment. We apply this model to a large corpus of online provider reviews. We show how the proposed model can leverage a small amount of data annotated by experts to guide topic/sentiment discovery. This extends our earlier work in which we introduced the probabilistic machinery leveraged here. In this communication we present a novel empirical evaluation of this model over an expanded corpus comprising nearly 60 000 physician reviews.

RELATED WORK
There has been a flurry of recent research concerning online physician-rating websites. Most related to the present work, Brody and Elhadad explored ‘salient aspects’ in online reviews of healthcare providers using latent Dirichlet allocation (LDA). Their approach was unsupervised, and did not use expert annotations. By contrast, we guide topic/sentiment discovery by leveraging a set of manually annotated reviews from a qualitative analysis, effectively combining qualitative and quantitative approaches.

DATA
RateMDs (http://ratemds.com) is a platform for patients to review doctors across four dimensions of care: helpful, knowledge, staff, and punctual. These are scored on a Likert scale of 1 (low) to 5.

RateMDs follows a URL structure that nests doctors alphabetically within states. Thus, with the aim of collecting a geographically diverse set of reviews, we sampled reviews as follows. We drew a state and a letter (A–Z), both uniformly at random. These two variables uniquely specify a page of RateMDs reviews, which we then downloaded. In this way we sampled 58 110 reviews of 19 636 unique US doctors. The median word count of sampled reviews is 41. Average scores (and SDs) for helpful, knowledge, staff, and punctual are 3.73 (1.71), 3.89 (1.60), 3.82 (1.50), and 3.73 (1.48), respectively. We show histograms of review scores across the four RateMDs dimensions in figure 1. We have made this corpus publically available (http://www.cebm.brown.edu/static/dr-sentiment.zip).

METHODS AND ANALYSIS
We leverage a probabilistic model based on factorial LDA (f-LDA) that captures both the sentiment and aspects latent in the free text of online provider reviews. The model accepts as input RateMDs reviews and infers from these the probabilable aspect of care (eg, interpersonal manner) and sentiment thereabout corresponding to every word in each review. To guide topic discovery, we use a small set of manual annotations created for a previously conducted qualitative study of online provider reviews’ via a method summarized below and described in detail elsewhere.
In previous qualitative work, López et al.⁶ identified the following important facets of online physician reviews: interpersonal manner, technical competence, and systems issues. We show examples in Table 1. These aspects were generated using inductive qualitative analysis informed by grounded theory,¹⁸ and are therefore more likely to capture meaningful content of online reviews than the categories imposed by RateMDs. Thus we would like a model that uncovers sentiment across these aspects in each review. We also want to exploit the available RateMDs ratings (which are close to, but not the same as, the target aspects). We thus defined a mapping from the RateMDs aspects to those defined by López et al. (see online supplementary appendix table S2; we ignore punctuality because it did not map onto the target aspects).

**Capturing aspect and sentiment using factorial LDA**

LDA¹⁵ is a generative model of text that assumes words in a document reflect a mixture of latent topics (each word is associated with a single topic). Topics index into distributions over words. Factorial-LDA (f-LDA)¹⁶ generalizes LDA to allow each token to be associated with a vector of latent topics, rather than only one. Here we consider a two-dimensional model in which each token is associated with one variable dictating its aspect and the other its sentiment.

f-LDA thus allows us to associate each review with a joint distribution over aspects and sentiment. Furthermore, f-LDA allows us to place rich prior distributions over model parameters. This provides the machinery to incorporate prior information into the model, including (1) data manually labeled with aspect and sentiment information by domain experts, and (2) user ratings included in the RateMDs data. Thus we can guide the model to uncover specific aspects of interest by leveraging this side information described above through the priors.

For additional technical details regarding the model, we refer the reader to our previous work⁵ and to the online supplementary appendix.

**EXPERIMENTAL RESULTS**

In preliminary work we showed that the f-LDA model can predict the user ratings in reviews with lower error than baseline ‘bag-of-words’ or LDA approaches.⁵ This suggests that our model is learning salient characteristics of the text. Here we

---

**Table 1  Annotations from López et al.⁶**

<table>
<thead>
<tr>
<th>Systems</th>
<th>Negative</th>
<th>Technical</th>
<th>Negative</th>
<th>Interpersonal</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly staff, short waits,</td>
<td>Difficult to park, rude staff, expensive</td>
<td>Good decision maker, knowledgeable</td>
<td>Poor decision maker</td>
<td>Empathic, communicates well</td>
<td>Poor listener, judgmental</td>
</tr>
<tr>
<td>convenient location</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
perform an in-depth analysis of the model output and evaluate it against ground-truth data.

We explore US state-level associations between external state-level healthcare statistics (percentage of patients who saw their primary care physician (PCP) within 14 days of discharge, mortality rates, and mean monetary expenditure, taken from the Dartmouth Atlas of Healthcare\(^\text{19}\)) and the model-inferred (latent) topic and sentiment prevalence in reviews using the hierarchical regression described below. In brief, we compared the fit of regressions using versus not using the information generated by the f-LDA model using likelihood ratio (LR) tests. If adding f-LDA model output results in statistically significantly

<table>
<thead>
<tr>
<th>Table 2 Highest ranking (most probable) words for each aspect and polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Systems</strong></td>
</tr>
</tbody>
</table>
|**Loves** | Loves | Charged | Son | MRI | Excellent | Arrogant |**Kids** | Kids | Pharmacy | Charged | Son | MRI | Excellent | Arrogant |**Awesome** | Awesome | Told | Gyn | Foot | Notch | Report |**Wonderful** | Wonderful | Awful | Delivered | Bleeding | Caring | Drug |**Love** | Love | Unprofessional | Happy | Breast | Ray | Compassionate | Misdiagnosed |**Loved** | Loved | Delivery | Hurt | Exceptional | Prescribed |**Comfortable** | Comfortable | Terrible | Ob | Bone | Best | License |**Knowledgeable** | Knowledgeable | Billed | pregnancies | Painful | Knowledgeable | Lack |**Explanations** | Explanations | Rude | Baby | Remove | Outstanding | Drugs | Dentist | Dentist | Records | Obgyn | Dentist | Little | Meds |**Pleasure** | Pleasure | Unhelpful | Saved | Cost | Thoughtful | Accused |**Informative** | Informative | Cancel | Pregnancies | Cost | Provides | Dissuasive |**Pediatrician** | Pediatrician | Refill | Decision | Cause | Considerate | Prescribe |**Highly** | Highly | Consultation | Amazing | Scan | Painful | Eventually | Effects |**Children** | Children | Double | Happier | Xrays | Dedicated | Effects |**Great** | Great | Paper | Team | Caused | Cause | Truly | Blood |**Smile** | Smile | Prescription | Wonderful | Injury | Reservation | Dangerous |**Ease** | Ease | Requested | Tuck | Injection | Intelligent | Insisted |**Understood** | Understood | Forgot | Outcome | Cause | Cause |**Easy** | Easy | Company | Team | Caused | Cause |**Knowledgeable** | Knowledgeable | Yelled | Deliver | Mouth | Amazing | Beware |**Fantastic** | Fantastic | Unacceptable | Thank | Xray | Hesitate | Poor |**Gentle** | Gentle | Sorry | Choose | Confirmed | Confused |**Personable** | Personable | Beware | Greatest | Mess | Genuinely | Addict |**Friendly** | Friendly | Said | Yous | Fix |”

better fitting models, it indicates that this model output contains information not readily available from the raw data.

We regress each state-level healthcare statistic against the state-level average ratings across the aforementioned four RateMDs categories (regression a). We then add as predictors variables corresponding to the mean overall frequency of inferred aspect and sentiment categorizations of each word in each review from the f-LDA model (regression b). These averages are calculated for a specific state by sampling from the f-LDA model for each token in each review for the said state (see online supplementary appendix table S3 for review counts). Specifically, we sample every word in every review for each state from the model posterior 100 times and calculating the average frequency with which words are assigned to each aspect/sentiment tuple. This results in 3 aspects×2 polarities=6 attributes per state. For example, one such attribute corresponds to the fraction of words in a given US state that the model assigned to the interpersonal/negative aspect/sentiment pair. We append these topic modeling output terms to the baseline average RateMDs ratings to realize model b (a is nested within b).

Denoting the outcome for state i by \( y_i \), the predictive attributes for state i by \( x_i \) (either regression a or b) and a heteroscedastic noise term for state i by \( e_i \), we assume:

\[
y_i = \beta_0 + \beta_i + \beta x_i + e_i
\]

where \( \beta_0 \) is an overall intercept and \( \beta_i \) is a zero-centered intercept for state i with between states-variance \( \tau^2 \):

\[
\beta_i \sim N(0, \tau^2)
\]

We define the per-state residual:

\[
e_i \sim N(0, \sigma^2)
\]

For the health outcomes \( y_i \), we first consider two statistics from the Dartmouth Atlas of Health Care\(^{20}\): the percentage of patients who visited a PCP within 14 days of hospital discharge following an acute event in 2010,\(^{19}\) and overall Medicare state mortality rates from 2007. (For both metrics we used the most recent available data.)

We found evidence for association between positive sentiment, based on the variables constituting both models, and the percentage of patients who saw their PCP within 14 days of discharge (p=0.03 regression b). This is a measure of adequate healthcare access and coordination of care. Furthermore, regression b (which includes f-LDA output) seems to explain more of the variance in this outcome than the RateMDs ratings alone: LR test p=0.10; \( R^2 \) of 0.13 when using RateMDs ratings only and 0.21 when including model output (\( R^2 \) is a measure of model fit). No association with individual predictors or difference between the fit of regressions a and b is seen for mortality, in line with expectations. It would indeed be surprising if online ratings tracked mortality rates: online ratings are an approximation of patient satisfaction, and across multiple measures of patient satisfaction—even with rigorous population sampling—there is no consistent association with mortality.\(^{21–23}\)

We also considered the cost of care across states, in terms of healthcare expenditure per capita.\(^{24}\) We again find that including the topic modeling output explains more variance in the outcome across states than the RateMDs ratings alone (LR test p=0.02). Including topic modeling output (regression b) results in an R-squared of 0.25 with respect to cost while using only the RateMDs ratings (regression a) results in an \( R^2 \) of 0.03.

Online doctor review positive sentiment across states is thus associated with patient likelihood of receiving and attending a post-hospitalization appointment with his or her PCP and (weakly) with higher cost of care. Moreover, the text of the reviews, modeled as aspect and sentiment categories, contains information beyond the user ratings that have been considered in previous studies.\(^{7}\) However, we emphasize that these are ecological associations, that is, the populations of patients who wrote the reviews are not the same populations in which outcomes were measured. Nonetheless, that inclusion of topic

Figure 2  Relative frequencies of target aspects over states.
modeling output better explains exogenous healthcare measures suggests that the proposed model recovers useful information otherwise latent in the review texts.

EXPLORATORY ANALYSES

In table 2 we reproduce the top-ranking (highest probability) words across each aspect/sentiment pair. Positive words tend to reflect general positive sentiment. By contrast, the negative words are more concrete, suggesting that negative reviews discuss specific healthcare experiences. This is consistent with prior research that has shown that dissatisfaction is not merely the absence of satisfaction, but a separate sentiment.

Figure 2 displays relative frequencies of aspects across states, illustrating the relative importance of different aspects geographically and perhaps reflecting differing local expectations for healthcare. Figure 3 shows the (marginal) state-level sentiment inferred in reviews from different states. It is unsurprising that sentiment varies given well-described geographic variation in healthcare delivery.

CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

We have proposed an f-LDA-based generative model of text to recover sentiment across different aspects of care latent in online reviews of physicians. This model leveraged existing, qualitatively annotated data. We showed that including f-LDA output in regression models improves correlations with state-level health outcome measures. This work demonstrates the potential of combining traditional qualitative analysis with large-scale quantitative modeling to facilitate analysis of online physician reviews.

This work has several limitations. RateMDs is one of many websites with patient rating data. We did not distinguish among types of medical care for which drivers of positive sentiment are likely to differ. Finally, the reported associations are ecological in nature.

Our results have several implications. First, traditional qualitative analysis can inform and enhance large-scale computational approaches to text data. Second, online doctor reviews correlate geographically with measures of healthcare coordination and quality. Finally, our results may suggest that higher patient satisfaction correlates with higher costs of care. This agrees with prior studies that suggest Americans seem to equate more medical care with higher-quality care.

Contributors The study was conceived by all authors. BCW wrote code to collect and process physician ratings. MJP and MD wrote the topic modeling code. BCW and TAT performed statistical analyses. SU provided qualitative ratings of online reviews and qualitative analysis of results/model output. BCW wrote first draft of manuscript; all authors edited.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement We will share all ~60 000 reviews analyzed in this article freely.

REFERENCES


18 Corbin J, Strauss A. Basics of qualitative research: techniques and procedures for developing grounded theory. Sage, 2008.
23 Corbin J, Strauss A. Basics of qualitative research: techniques and procedures for developing grounded theory. Sage, 2008.