A Large-Scale Quantitative Analysis of Latent Factors and Sentiment in Online Doctor Reviews

Byron C. Wallace, Brown University, Dept of Health Services Policy & Practice, Box G-S121-8, Providence, RI 02912, USA, (401)-863-6421, byron_wallace@brown.edu

Michael J. Paul, Johns Hopkins University, Dept of Computer Science, Baltimore, Maryland, USA

Urmimala Sarkar, UCSF, Dept of Medicine, San Francisco, California, USA

Thomas A. Trikalinos, Brown University, Dept of Health Services Policy & Practice, Providence, RI, USA

Mark Dredze, Johns Hopkins University, Human Language Technology Center of Excellence, Baltimore, Maryland, USA

Keywords: social media, natural language processing, physician reviews, topic modeling

Word count: 1967

A Large-Scale Quantitative Analysis of Latent Factors and Sentiment in Online Doctor Reviews

Abstract

Objective: 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.
Methods: We describe a probabilistic generative model that captures latent sentiment across aspects of care (e.g., *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.
Results: We report both qualitative and quantitative results. Model output correlates with state-level measures of quality healthcare, including patient likelihood of visiting their PCP within fourteen days of discharge (p=.03), and using the proposed model better predicts this outcome (p=.10). We find similar results for healthcare expenditure.
Conclusions: 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 [1] 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 [2], a recent study of a random sample of 500 urologists found online reviews for about 80% of them [3].

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 [4]. 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 [5] 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 [3 6-13]. Most related to the present work, Brody and Elhadad [14] explored "salient aspects" in online reviews of healthcare providers using Latent Dirichlet Allocation (LDA) [15]. Their approach was unsupervised, 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.

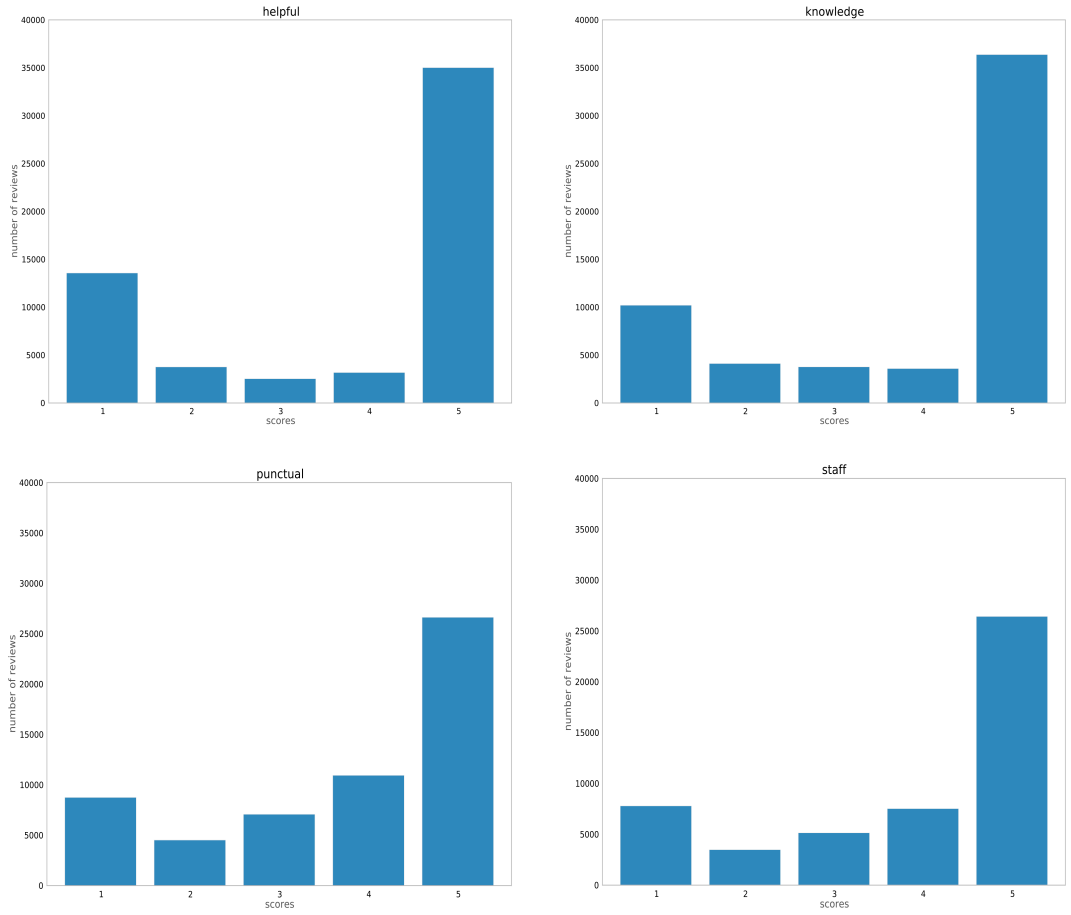


Figure 1. Histograms of observed scores across RateMDs data with respect to the aspects defined by RateMDs (clockwise from top left: *helpful, knowledge, staff, punctual)*.

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 standard deviations) 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 Latent Dirichlet Allocation (f-LDA) [16] that captures both the sentiment and aspects latent in the free text of online provider reviews [5]. The model accepts as input RateMDs reviews and infers from these the probable aspect of care (e.g., 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 [6] via a method summarized below and described in detail elsewhere [5 17].

Table 1. Annotations from [6].

Systems		*Technical*		*Interpersonal*	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
friendly staff, short waits, convenient location	difficult to park, rude staff, expensive	good decision maker, knowledgeable	poor decision maker	empathic, communicates well	poor listener, judgmental

In previous qualitative work, López et al. identified the following important facets of online physician reviews: *interpersonal manner, technical competence* and *systems issues* [6]. 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. (Appendix Table 2; we ignore *punctuality* because it did not map onto the target aspects).

Capturing Aspect and Sentiment Using Factorial LDA

LDA [15] 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.*

Factorial-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 work [5] and to the 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 perform an in-depth analysis of the model output and evaluate it against ground-truth data.

We explore U.S. state-level associations between external state-level health care statistics (percentage of patients who saw their PCP within fourteen days of discharge, mortality rates and mean monetary expenditure, taken from the Dartmouth Atlas of Healthcare [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 better fitting models, it indicates that this model output contains information not readily available from the raw data.

We regress each state-level health-care statistic against the state-level average ratings across 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 said state (see Appendix Table 3 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 U.S. 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 $\boldsymbol{x}_i$ (either regression *a* or *b*) and a heteroscedastic noise term for state *i* by e_i, we assume:

$$y_i = \beta_0 + \beta_i + \boldsymbol{\beta}\boldsymbol{x}_i + e_i$$

where β_0 is an overall intercept and β_i is a zero-centered intercept for state *i* with between states-variance τ^2:

$$\beta_i \sim N(0, \tau^2)$$

We define the per-state residual:

$$e_i \sim N(0, s_i^2)$$

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 fourteen 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 fourteen days of discharge (p=.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=.10; R-squared of .13 when using RateMD ratings only and .21 when including model output). 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 health care 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=.02). Including topic modeling output (regression b) results in an R-squared of .25 with respect to cost while using only the RateMDs ratings (regression a) results in an R-squared of .03.

Online doctor review positive sentiment across states thus is 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, i.e., the populations of patients who wrote the reviews are not the same populations in which outcomes were measured. Nonetheless, that inclusion of topic modeling output better explains exogenous health care measures suggests that the proposed model recovers useful information otherwise latent in the review texts.

Table 2. Highest ranking (most probable) words for each aspect and polarity.

Systems		*Technical*		*Interpersonal*	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
loves	charged	son	mri	excellent	arrogant
kids	pharmacy	gyn	foot	notch	report
awesome	told	delivered	bleeding	caring	drug
wonderful	awful	breast	ray	compassionate	misdiagnosed
love	unprofessional	thankful	nerve	highly	reaction
loved	paying	delivery	hurt	exceptional	prescribed
comfortable	terrible	ob	bone	best	license
knowledgable	billed	children	antibiotic	knowledgeable	lack
explains	rude	baby	remove	outstanding	drugs
dentist	records	obgyn	dentist	wonderful	meds
sweet	refused	pregnancies	painful	honest	dismissed
pleased	unhelpful	saved	cost	thoughtful	accused
informative	cancel	pregnancy	crying	provides	dismissive
pediatrician	refill	section	teeth	genuine	ordered
highly	consultation	decision	causing	considerate	prescribe
children	double	amazing	scan	pleasure	eventually
great	paper	happier	xrays	dedicated	effects
smile	prescription	wonderful	injury	reservation	dangerous
ease	requested	team	caused	truly	blood
understood	forgot	outcome	cause	humor	basic
easy	company	tuck	injection	intelligent	insisted
knowledgeable	yelled	deliver	mouth	amazing	beware
fantastic	unacceptable	thank	xray	hesitate	poor
gentle	sorry	choose	confirmed	attentive	wrote
personable	beware	greatest	mess	genuinely	addict
friendly	said	youn	fix	insightful	jerk
calming	disrespectful	infertility	damage	listens	signs
prompt	apology	daughter	insisted	team	repeatedly
fabulous	worst	child	tooth	loving	refused
efficient	lunch	babies	needle	highest	uncaring
amazing	form	best	fusion	understanding	enemy
earth	covered	supportive	severe	knowledgable	records
caring	contact	pleased	arm	incredible	careless
adore	canceled	deliveries	canal	respectful	eat
helpful	letter	control	pulled	earth	ignored
knows	ended	handled	stated	mile	medication
understanding	horrible	talent	spinal	fantastic	reported
parents	refund	blessed	disc	thorough	incorrect
atmosphere	denied	boy	shots	talented	lose
attentive	cash	highly	said	skillful	behavior
equally	ridiculous	pregnant	cast	supportive	unsympathetic
helpfull	occasions	twins	herniated	explains	errors
comforting	cancelled	confident	refused	warm	pressure
pleasant	response	vegas	muscle	unique	incompetent
calm	charges	cardwell	infected	chiropractic	depressed
thorough	dirty	bless	infection	fortunate	avoid
warm	forced	miscarriages	dental	blessed	unprofessional
adults	brief	forward	throat	fabulous	board
nicest	disorganized	watabe	crown	respected	social
satisfied	money	skilled	telling	superb	insulted

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 [25 26].

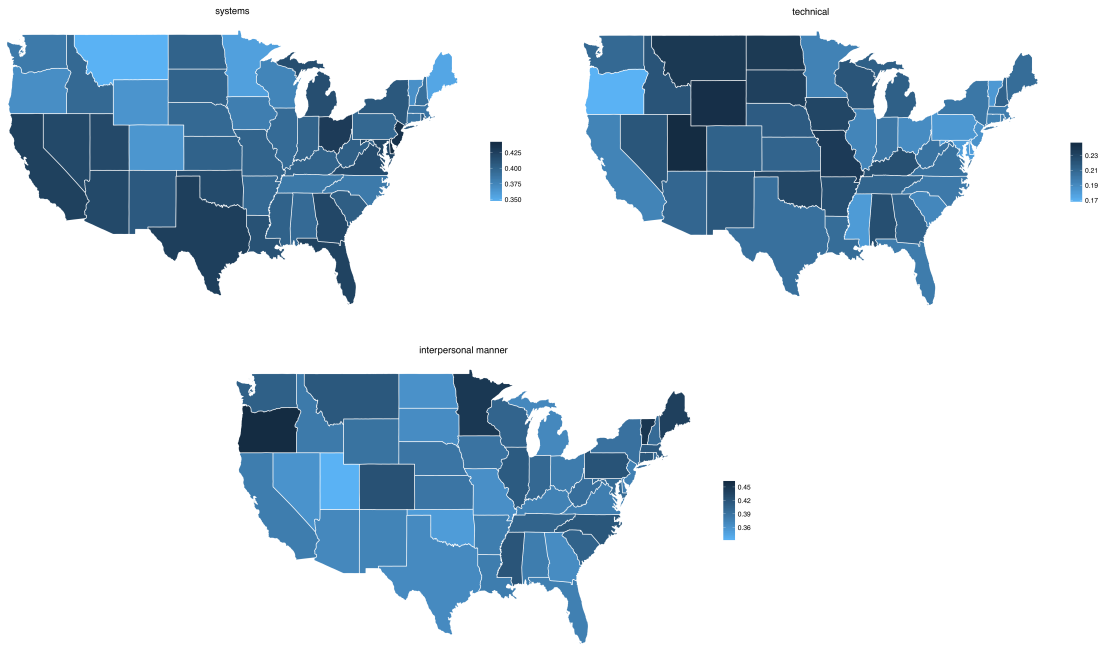


Figure 2. Relative frequencies of target aspects over states.

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 [20].

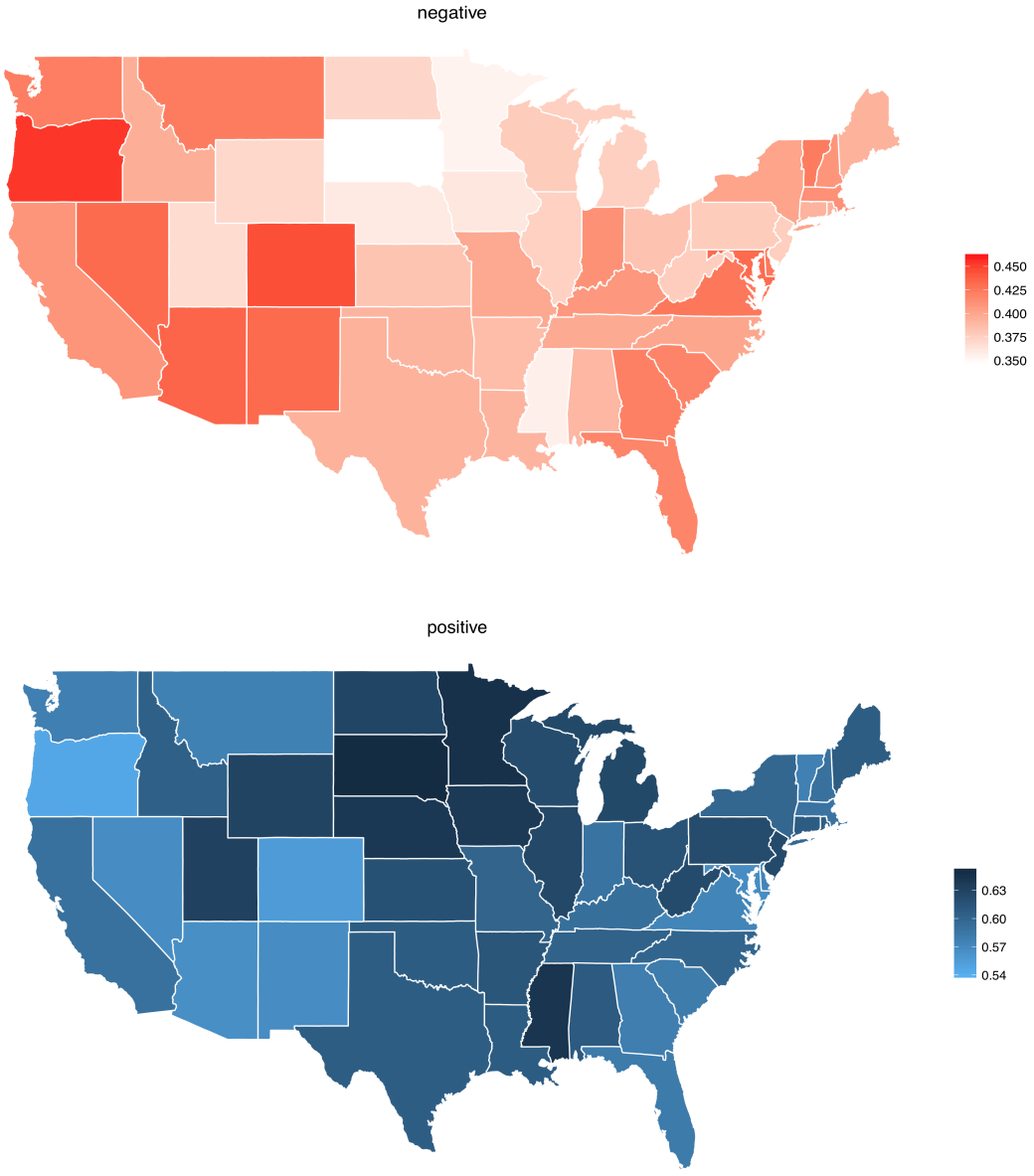


Figure 3. Relative frequencies of negative (top) and positive (bottom) sentiment (marginalized over aspects) across states. We show both for convenience; they are of course symmetric.

Conclusions, Limitations and Future Work

We have proposed a factorial Latent Dirichlet Allocation (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 key 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 [21 27].

Funding Statement

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Contributorship Statement

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 Statement

The authors have no competing interests to declare.

References

1. *Fox S, Duggan M. Health online 2013. Health 2013*
2. *Tara Lagu MD M, Kaufman EJ, Asch DA, et al. Content of weblogs written by health professionals. Journal of general internal medicine 2008;**23**(10):1642-46*
3. *Ellimoottil C, Hart A, Greco K, et al. Online reviews of 500 urologists. The Journal of urology 2012*
4. *Greaves F, Ramirez-Cano D, Millett C, et al. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ quality & safety 2013;**22**(3):251-55*
5. *What Affects Patient (Dis) satisfaction? Analyzing Online Doctor Ratings with a Joint Topic-Sentiment Model. AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI); 2013.*
6. *López A, Detz A, Ratanawongsa N, et al. What patients say about their doctors online: a qualitative content analysis. Journal of general internal medicine 2012;**27**(6):685-92*
7. *Segal J, Sacopulos M, Sheets V, et al. Online doctor reviews: do they track surgeon volume, a proxy for quality of care? Journal of medical Internet research 2012;**14**(2)*
8. *Emmert M, Sander U, Pisch F. Eight questions about physician-rating websites: A systematic review. Journal of medical Internet research 2013;**15**(2)*
9. *Galizzi MM, Miraldo M, Stavropoulou C, et al. Who is more likely to use doctor-rating websites, and why? A cross-sectional study in London. BMJ open 2012;**2**(6)*
10. *Greaves F, Pape UJ, King D, et al. Associations between web-based patient ratings and objective measures of hospital quality. Archives of internal medicine 2012;**172**(5):435-36*
11. *Alemi F, Torii M, Clementz L, et al. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. Quality Management in Healthcare 2012;**21**(1):9-19*
12. *Greaves F, Ramirez-Cano D, Millett C, et al. Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. Journal of medical Internet research 2013;**15**(11)*
13. *Lagu T, Goff SL, Hannon NS, et al. A mixed-methods analysis of patient reviews of hospital care in England: implications for public reporting of health care quality data in the United States. Joint Commission Journal on Quality and Patient Safety 2013;**39**(1):7-15*
14. *An unsupervised aspect-sentiment model for online reviews. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics; 2010. Association for Computational Linguistics.*
15. *Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. the Journal of machine Learning research 2003;**3**:993-1022*
16. *Factorial LDA: Sparse multi-dimensional text models. Advances in Neural Information Processing Systems 25; 2012.*

17. *Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. Proceedings of NAACL-HLT; 2013.*
18. *Corbin J, Strauss A. Basics of qualitative research: Techniques and procedures for developing grounded theory: Sage, 2008.*
19. *Goodman DC, Fisher ES, Chang C-H, et al. After hospitalization: a Dartmouth atlas report on post-acute care for Medicare beneficiaries. The Dartmouth Institute for Health Policy & Clinical Practice 2011:1-52*
20. *Goodman DC, Fisher ES, Wennberg J, et al. The Dartmouth Atlas of Health Care. Secondary The Dartmouth Atlas of Health Care 2013.* http://www.dartmouthatlas.org/.
21. *Fenton JJ, Jerant AF, Bertakis KD, et al. The cost of satisfaction: a national study of patient satisfaction, health care utilization, expenditures, and mortality. Archives of internal medicine 2012:archinternmed. 2011.1662 v1*
22. *Schneider EC, Zaslavsky AM, Landon BE, et al. National quality monitoring of Medicare health plans: the relationship between enrollees' reports and the quality of clinical care. Medical care 2001:1313-25*
23. *Odigie EG, Marshall R. Quality monitoring of physicians: linking patients' experiences of care to clinical quality and outcomes. Journal of general internal medicine 2008;**23**(11):1784-90*
24. *Health Expenditures by State of Residence. Secondary Health Expenditures by State of Residence 2011.* http://www.cms.gov/NationalHealthExpendData/downloads/resident-state-estimates.zip.
25. *Beck RS, Daughtridge R, Sloane PD. Physician-patient communication in the primary care office: a systematic review. The Journal of the American Board of Family Practice / American Board of Family Practice 2002;**15**(1):25-38*
26. *Anderson RT, Camacho FT, Balkrishnan R. Willing to wait?: The influence of patient wait time on satisfaction with primary care. BMC Health Services Research 2007;**7**(1):31*
27. *Lyles CR, López A, Pasick R, et al. "5 Mins of Uncomfyness Is Better than Dealing with Cancer 4 a Lifetime": an Exploratory Qualitative Analysis of Cervical and Breast Cancer Screening Dialogue on Twitter. Journal of Cancer Education 2012:1-7*

Appendix to *A Large-Scale Quantitative Analysis of Latent Factors and Sentiment in Online Doctor Reviews*

Seeding the Model with Labeled Data

As discussed in the article, we would like to guide topic discovery in light of the data manually annotated by López et al. [1] However this dataset is relatively small (843 reviews), while our corpus of unannotated reviews is very large (58,110 reviews). We thus adopt a semi-supervised strategy in which we leverage the manually annotated data to define priors over the f-LDA parameters, as has been proposed in previous work [2]. In particular, this involves inducing a supervised model over word probabilities, where we assume:

$$P(w|topic, sentiment) \propto \exp\left(\eta_w^b + \eta_w^{topic} + \eta_w^{sentiment}\right)$$

Appendix Equation 1. Log-linear component model of a word given associated topic and sentiment.

Thus defining independent components corresponding to the conditioning *topic* and *sentiment* values. This is a variant of the SAGE model proposed by Eisenstein et al. [3]. It has the same form as Appendix Appendix Equation 1, but assumes that every word comprising a given review shares the same topic and sentiment (rather than allowing words to associate with different topics and sentiment). We estimate the parameters of this model over the aforementioned small set of manually annotated data (via gradient ascent) and then use these estimates as the means of Gaussian priors over the ω_w^t terms defined above, e.g., $\omega_w^{topic} \sim Normal(\eta_w^{topic}, \sigma^2)$.

For the generative story of our model, see Appendix Table 1; we have also depicted the model graphically in Appendix Figure 1. The basic intuition is that we share parameters between word distributions for <*aspect*, *sentiment*> pairs that share components. Similarly, parameters governing the distribution over pairs within documents share components.

Intuitively, <*aspect*, *sentiment*> pairs that share a component (e.g., the same aspect but different sentiment) should exhibit similarities in their associated word distributions ϕ. To tie these components together, f-LDA defines the Dirichlet priors for these distributions such that they are shared between tuples *t* that share components. Specifically, word distributions ϕ^t for each tuple *t* has a prior Dir($\boldsymbol{\omega}^t$), where $\boldsymbol{\omega}^t$ has dimensionality equal to the vocabulary size *V*. We define the entry corresponding to word *w* in this vector as a log-linear function including components corresponding to *aspect* and *sentiment*:

$$\omega_w^t = \exp\left(\omega^b + \omega_w^0 + \omega_w^{aspect} + \omega_w^{sentiment}\right)$$

Where ω^b is a corpus-wide bias term, ω_i^0 is an intercept for word *i*, ω_i^{topic} and $\omega_i^{sentiment}$ are aspect- and sentiment-specific biases for word *i*, respectively. (Here we are assuming *aspect* and *sentiment* are those in the associated tuple *t*.) Thus each *aspect* and *sentiment* pair is associated with its own vector over the vocabulary, and a given tuple reflects both of its constituent components. We place independent Normal priors on all ω terms, with means η. The means η are given by first training a similar but fully supervised log-linear model on the data manually annotated by López et al. [1], and taking the parameters learned from this model as η. We provide

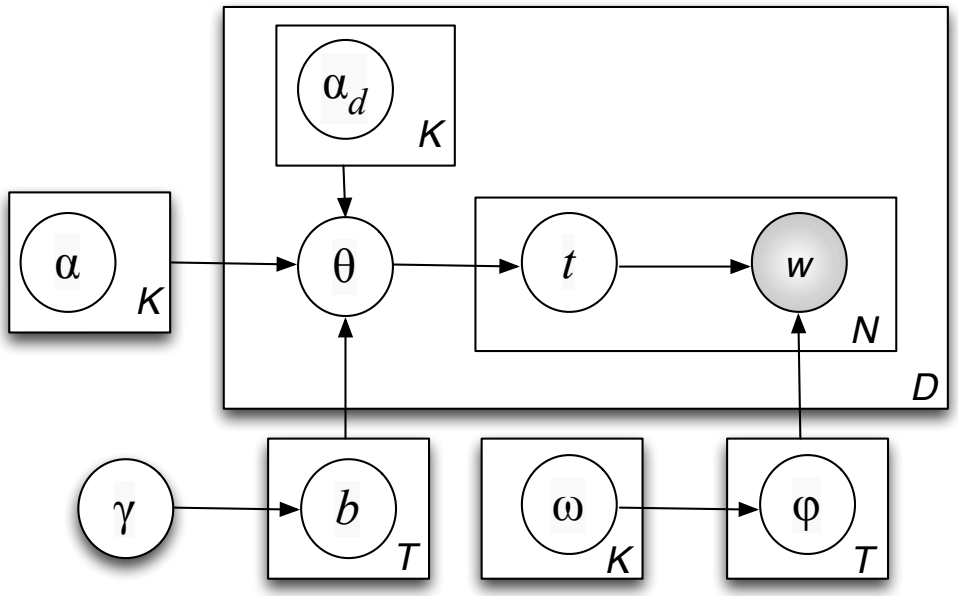
details of this procedure in the Appendix. We use this semi-supervised approach of incorporating supervision as *prior* knowledge, first used for f-LDA in [2], because the annotated dataset is much smaller than the full corpus.

We use a similar prior structure for the document mixture distributions θ so as to operationalize the intuition that, e.g., if *positive* sentiment is prevalent in a review, then it is also more likely to be prevalent across all aspects. We parameterize document mixture Dirichlet priors over tuples *t* with parameters α_t^d that reflect corpus- and review-specific biases. Specifically, we set these parameters as follows:

$$\alpha_t^d \propto \exp\left(\alpha^b + \alpha_t^{aspect} + \alpha_t^{aspect,d} + \alpha_t^{sentiment} + \alpha_t^{sentiment,d} + \rho\, r_t^d\right)$$

Where α^b is a corpus-wide bias parameter, α_t^{aspect} and $\alpha_t^{aspect,d}$ are corpus-wide and document-specific terms for topics, respectively, and we have included analogous terms for sentiment. The *r* variable represents the user rating (on a Likert scale from 1 to 5) for the corresponding aspect (Table 3), where the rating is differenced around the midpoint 3, and negated for the negative sentiment. This variable (scaled by ρ), is an additional extension to the standard f-LDA model which allows us to use the readily available review-level ratings (sentiment) across the aligned RateMDs aspects to inform the prior for each review. This approach suggests, e.g., the pair <*systems*,*positive*> should *a priori* have high probability if the user rating for *staff* is high.

We fit this model via a gradient ascent and collapsed Gibbs sampling strategy detailed elsewhere [4 5].



Appendix Figure 4. The graphical model of the f-LDA instantiation we used. There are *K*=5 components (*positive, negative*, *interpersonal, systems and technical*) and *T*=6 tuples thereof (e.g., <*technical*, *positive*>). The *b* variable is a sparsity inducing scalar that allows for specific tuples to have globally low probability. Together with global and review specific $\boldsymbol{\alpha}$ priors, this gives rise to a distribution $\boldsymbol{\theta}$ over tuples for each review; these tuples in turn generate the tokens (words) comprising reviews.

Generative story of the aspect-sentiment f-LDA model	
Input: 1. Observable document collection *D*, each document *d* is defined as word vector $\boldsymbol{w}^d$ 2. For each document *d*, a vector of user ratings $\boldsymbol{r}^d$ 3. Set of weight vectors $\boldsymbol{\eta}$ trained on annotated data, to be used as priors for the word distribution hyperparameters	
1. Draw hyper-parameters $\boldsymbol{\omega}$ from Normal($\boldsymbol{\eta}$, $\mathbf{I}\sigma^2$) 2. Draw hyper-parameters $\boldsymbol{\alpha}$ and ρ from Normal($\mathbf{0}$, $\mathbf{I}\sigma^2$) 3. For each <*aspect*, *sentiment*> tuple *t* a. Sample $\phi_t \sim$Dirichlet($\boldsymbol{\omega}^t$) 4. For each document (review) $d \in D$ a. Draw document weights for each component *k:* $\alpha_k^d \sim$ Normal($\mathbf{0}$, $\mathbf{I}\sigma^2$) b. Sample document distribution over tuples: $\theta^d \sim$ Dirichlet(α^d) c. For each token *w*: i. Sample <*aspect, sentiment*> tuple $t \sim \theta^d$ ii. Sample word $w \sim \phi_t$ Where we define: $\omega_w^t = \exp(\omega^b + \omega_w^0 + \omega_w^{aspect} + \omega_w^{sentiment})$ and $\alpha_t^d = \exp(\alpha^b + \alpha_t^{aspect} + \alpha_t^{aspect,d} + \alpha_t^{sentiment} + \alpha_t^{sentiment,d} + \rho\, r_t^d)$	Parameters for word priors Parameters for doc. priors Word distributions Parameters for doc. priors Document distribution Latent variables Observed variable Word prior is function of general, aspect- and sentiment-specific weights Document prior combines aspect- and sentiment-specific weights and user rating *r*

Appendix Table 1. Generative story of the f-LDA variant we use.

RateMDs	López et al.
Knowledgeability	*Technical*
Staff	*Systems*
Helpfulness	*Interpersonal*

Appendix Table 2. Mapping from RateMDs tags to those specified by López et al.

State	Number of reviews
Alabama	1584
Alaska	797
Arizona	2344
Arkansas	635
California	1802
Colorado	737
Connecticut	1622
DC	865
Delaware	1330
Florida	1915
Georgia	1136
Hawaii	499
Idaho	1131
Illinois	675
Indiana	1042
Iowa	751
Kansas	1178
Kentucky	533
Louisiana	1301
Maine	739
Maryland	601
Massachusetts	925
Michigan	2033
Minnesota	1969
Mississippi	545
Missouri	576
Montana	583
Nebraska	1256
Nevada	2162
New Hampshire	826
New Jersey	2353
New Mexico	170
New York	2012
North Carolina	790
North Dakota	433
Ohio	772
Oklahoma	1448
Oregon	664
Pennsylvania	550
Rhode Island	1558
South Carolina	1607
South Dakota	760
Tennessee	957
Texas	1535
Utah	1247
Vermont	240
Virginia	2202
Washington	1590
West Virginia	1048
Wisconsin	1877

Appendix Table 3. Review counts per state in our corpus. Note that while we select states uniformly at random, we then select a page for said state based on surname first letter, also uniformly at random. We then download all of the reviews for this state / letter pair. Thus if a state is 'more popular' on RateMDs, it will likely have more reviews for any given letter (e.g., more reviews for physicians whose last name starts with 'A'). This is why we still have variable numbers of reviews per state.

Appendix References

1. López A, Detz A, Ratanawongsa N, et al. What patients say about their doctors online: a qualitative content analysis. Journal of general internal medicine 2012;**27**(6):685-92
2. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. Proceedings of NAACL-HLT; 2013.
3. Sparse additive generative models of text. Proceedings of the 28th International Conference on Machine Learning (ICML-11); 2011.
4. Factorial LDA: Sparse multi-dimensional text models. Advances in Neural Information Processing Systems 25; 2012.
5. What Affects Patient (Dis) satisfaction? Analyzing Online Doctor Ratings with a Joint Topic-Sentiment Model. AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI); 2013.