Detecting Salient Aspects in Online Reviews of Health Providers

Samuel Brody, Ph.D., Noémie Elhadad, Ph.D.
Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

We present a fully automated method to capture what topics health consumers discuss when reviewing their health providers online. Our method does not rely on any manual tagging of the information, and operates on the text of online reviews. We analyze a large set of reviews and compare the topics discussed when reviewing providers with different specialties. This work provides a complementary view on the traditional qualitative approaches proposed so far to capturing factors for patient satisfaction. Furthermore, our research contributes to understanding in a bottom-up fashion the needs and interests of health consumers online.

Introduction

As more and more individuals rely on the Internet as a source of information and decision aid for their health needs [1], there is a growing demand for automated tools, which can support the needs of health consumers online. One of the revolutions brought on by Web 2.0 technology is the ability for Web users to rely on each other's opinions when making decisions ranging from choosing a restaurant, renting a movie, to buying a laptop. This trend is now reaching the health domain, with a growing number of websites devoted to reviews of health practitioners authored by health consumers.

In a patient-centric practice, physicians have interest in understanding what matters to their patients when choosing a health provider. Patients on their side would benefit from understanding what aspects of a practice other patients pay attention to when choosing a provider. For health researchers, it is essential to analyze what factors health consumers care about when assessing a provider, as it can influence health communication strategies. Finally, from a consumer health informatics standpoint, providing tools to process and organize the information conveyed in provider reviews can augment the functionalities of Personal Health Records, provided the tools are accurate enough.

The traditional method to identify factors for patient satisfaction has been through surveys and questionnaires of patients, either to assess the effect of a particular element of the patient-provider interaction (e.g., [2]) or as a comprehensive analysis tool [3, 4]. But these factors are often established in a top-down fashion, from experts. Furthermore, because they are discussed in reports and papers for the scientific community, health consumers do not always rely on them when choosing providers. We propose the use of computational methods to conduct a complementary type of analysis: *discovering in a dynamic, bottom-up fashion the factors contributing to patient satisfaction through the quantitative analysis of the provider reviews authored online by health consumers.* By relying on the *collective experience* of health consumers as conveyed in the text of the reviews, we propose to identify the salient aspects about a provider that matter to the health consumers themselves.

While there has been much debate over the quality and impact of such source of information recently [5, 6, 7] (in particular the fear of fraudulent reviews and lack of trust in the authors), and much care has to be put into interpreting and using the results of any fully automated method of analysis, the phenomenon of peer reviewing seems to be a growing trend and a medium health consumers rely on more and more (if only measured as the ever-increasing number of reviews written by health consumers online and websites dedicated to this type of content) [8]. This particular work relies on the reviews to identify trends through the text mining of large amounts of reviews, thereby minimizing the impact of deceitful reviews.

Some websites provide a structured questionnaire for health consumers to review a provider. For instance, the website HealthGrades[1] allows for nine dimensions to be assessed, each on a 5-level scale: general recommendation (would you recommend this physician to friends and family), level of trust (do you trust the physician to make recommendations that are in your best interest), to which extent the physician helps patients understand their condition, to which extent the physician listens and answers questions, the time spent with patient, the ease of scheduling, the office environment (cleanliness, etc.), the friendliness of the office staff, and finally the wait time. Other websites provide a hybrid of structured questions and free-text for reviewers to enter. The websites RateMDs[2] and ZocDoc[3], for instance, provide ratable dimensions (ZocDoc lists three dimensions: overall recommendation, bedside manner and wait time, while RateMDs lists four: punctuality, helpfulness, knowledge and overall recommendation) but also allow users to enter their own review. The variation over websites indicate that provider reviews is still an emerging genre of texts, with no set standard for health consumers to follow.

[1] `www.healthgrades.com`
[2] `www.ratemds.com`
[3] `www.zocdoc.com`

The fact that the genre is still fluid is advantageous for a quantitative, bottom-up analysis, as our goal is to discover salient points of discussion in reviews, without being influenced by a particular website's organization of information.

Researchers in computational linguistics and information retrieval have investigated how to identify aspects and sentiment from text automatically (see [9] for a complete review of techniques). However, most work to date has focused on product reviews (e.g., laptops, restaurants, movies). Applying computational methods to the analysis of reviews of health providers is timely and novel.

The main research questions we focus on for this study are: (i) what are the salient topics or aspects discussed by health consumers when reviewing health providers? (ii) can such aspects be discovered automatically in a bottom-up fashion from the text of the reviews? and (iii) to which extent are the aspects specific to the providers' specialties?

Methods

We describe the main computational method on which we rely (Latent Dirichlet Analysis, or LDA) to identify salient aspects in reviews of health providers and how we customize it to answer our research questions. There are two challenges we address in particular: (i) dataset selection and selection of the unit of processing on which to apply LDA, and (ii) determining the optimal number of aspects discussed in the reviews (model order). We first give an overview of LDA in general, followed by our experimental setup.

Latent Dirichlet Allocation. Our method for determining common topics discussed in medical reviews is based on a generative probabilistic graphical model, Latent Dirichlet Allocation (LDA) [10]. LDA is a fully unsupervised method to identify common topics of discussion in a collection of documents. The topics are identified automatically, without requiring any prior knowledge or manual annotation. This is particularly attractive to our task, since we want to *discover* the common topics discussed in reviews of health providers, rather than making hypotheses about the aspects of a health provider practice that are important to health consumers and validating them through data analysis.

LDA is a generative probabilistic model of a text collection. Documents (in our study, individual reviews) are represented as random mixtures over latent topic, where each topic is characterized by a distribution over words. The distribution of topics is assumed to come from a family of parametrized Dirichlet distributions. The words in a document are generated one after the other by repeatedly samplic a topic according to the topic distribution and selecting a word given the chosen topic.

The main utility of the model is achieved by reversing it and inferring from a collection of preexisting documents both the unknown topics and the words with which they are associated. In the inference procedure the most likely topic distribution and word assignments are calculated from the observed data. The output of the inference procedure is a list of topics and the probabilities for each word in the data appearing in that topic. Since the model is unsupervised, it does not provide labels (names) for the topics. By examining the most probable words for each topic (see Figure 1), we can get an idea of its subject, and a label can be assigned. The generative nature of the model allows it to handle newly observed documents which do not conform precisely to a previously seen distribution.

The authors of [10] compare LDA to other models proposed in the literature, and report improved results on document modeling and text classification tasks, where their model does considerably less over-fitting than the others. Since then, LDA has been applied to many tasks, such as entity resolution [11], information retrieval [12] and image processing [13]. Several efficient methods have been developed for inference with LDA. In this work, we employ a standard implementation of LDA which uses Gibbs sampling for parameter estimation and Inference.[4]

Data. In order to create our datasets, we collected a corpus of reviews from the public RateMDs website. As a preprocessing step, the portions containing the reviews were extracted from the HTML pages, along with the specialty designation of each provider. The reviews were tokenized and separated to individual sentences. Stop words were removed. We stratified our dataset of reviews into four individual sets of reviews: review of general practitioners (GP), obstetricians/gynecologists (ObGyn), dentists (Dent), and psychiatrists (Psych).

LDA for Reviews of Health Providers. A specially-tailored model [14], based on LDA, was shown to be effective at finding rateable aspects of hotel reviews, with the help of additional aspect-specific information provided by the reviewers. In [15], the authors demonstrated that a local version of LDA, which operates on individual sentences rather than documents, and doesn't require additional information, can find rateable aspects in a variety of domains including product and restaurant reviews. We hypothesize that a similar approach would be suitable for the domain of professional services and, in particular, for our task of determining the salient aspects in online reviews of health providers.

Model Order The issue of model order, i.e., determining the correct number of clusters (in our case the

[4] GibbsLDA++, by Xuan-Hieu Phan. Available at `http://gibbslda.sourceforge.net/`.

discovered topics), is an important element in unsupervised learning. A common approach [16, 17] is to rely on a cluster validation procedure. In such a procedure, different model orders are compared, and the one with the most consistent clustering is chosen. For the purpose of the validation procedure, we have a cluster corresponding to each aspect, and we label each sentence as belonging to the cluster of the most probable aspect.

Given the collection of sentences in our data, D, and two connectivity matrices C and $\hat{C}$, where a cell i, j contains 1 if sentences d_i and d_j belong to the same cluster, we define a consistency function F (following [17]):

$$F(C, \hat{C}) = \frac{\sum_{i,j} 1\{C_{i,j} = \hat{C}_{i,j} = 1, d_i, d_j \in \hat{D}\}}{\sum_{i,j} 1\{C_{i,j} = 1, d_i, d_j \in \hat{D}\}} \quad (1)$$

We then employ the following procedure:

1. Run the LDA model with k topics on D to obtain connectivity matrix C_k.
2. Create a comparison connectivity matrix R_k based on uniformly drawn random assignments of the instances.
3. Sample random subset D^i of size $\delta|D|$ from D.
4. Run the LDA model on D^i to obtain connectivity matrix C_k^i.
5. Create a comparison matrix R_k^i based on uniformly drawn random assignments of the instances in D^i.
6. Calculate $score_i(k) = F(C_k^i, C_k) - F(R_k^i, R_k)$ where F is given in Eq. 1.
7. Repeat steps 3 to 6 q times.
8. Return the average score over q iterations.

This procedure calculates the consistency of our clustering solution, using a similar sized random assignment for comparison. It does this on q subsets to reduce the effects of chance. The k with the highest score is chosen. In our experiments, we used $q = 10, \delta = 0.9$, and let k range from four to fifteen.

Results

The corpus contained 33,654 reviews of 12,898 medical practitioners in the NY state area. The reviews are often quite short, with an average of 4.17 sentences per review, and 15.5% of them containing only one line. The breakup of medical specialties in the data is given in Table 1. The overall dataset contains reviews about 42 additional specialties (ranging from ophthalmologists to chiropractors to anesthesiologists).

Our cluster-validation scheme for determining model order detected the most consistent set of aspects to contain six for the G.P. and OB/Gyn data, and four and five aspects for the dentist and psychiatrist reviews, respectively. The automatically discovered aspects, along with an example sentence for each, are given in Figure 1 (the labels are not an output of the method, and were provided by the authors).

Specialty	# Reviews
Internist	1,777
Gynecologist	1,389
Family / General	1,296
Pediatrist	789
Dentist	774
Psychiatrist	559
Orthopedist	559
Cardiologist	471
Gastroenterologist	428
Dermatologist	426
All Specialties	33,654

Table 1. Breakup of our online review dataset, listing the number of reviews for each of the top ten specialties, and the overall total number of reviews.

Aspect	GP	Dent	ObGyn	Psych
Recommend.	✓	✓	✓	✓
Manner	✓	✓	✓	✓
Anecdotal	✓	✓	✓	✓
Attention	✓	–	✓	✓
Scheduling	–	–	✓	✓
Special	Prescrip. & Tests	Cost	Pregnancy	–

Table 2. Summary showing the shared and distinct aspects in the datasets.

When examining the inferred aspects for the different specialties, we can see that there are several aspects which are shared between many specialties (though the details vary between them). There are also aspects that are specific to one or two specialties (or not strong enough in the others to merit a separate aspect). Table 2 summarizes these findings.

Discussion

In the effort of discovering what health consumers consider salient aspects when reviewing providers, we had a set of desiderata for our computational methods: dynamic and bottom-up, without any reliance on manual annotation. Our results show that LDA is an appropriate method given our constraints. Furthermore, when reviews are processed at the sentence level (rather than as a whole), and reviews are grouped by specialty, it is possible to identify salient aspects that are specialty-specific.

The discovered aspects which are common to all specialties resemble the traditional aspects of patient satisfaction questionnaires (such as bedside manner of the staff and the provider, and level of attention provided by the provider to the patients). When examining the aspects that are specific to different specialties, however, interesting patterns emerge. For instance, cost is a salient topic only for dentists. This makes sense,

Family/General Practitioner
1. **Prescrip. & Tests:** call, results, blood, test, pain *“[...] brand name prescription instead of the many generic ones available so as to make commissions [...]”*
2. **Manner:** staff, great, caring, knowledgeable *“Terrific doctor, excellent bed-side manner, staff needs to improve [...]”*
3. **Anecdotal:** visit, problem, insurance, treatment *“his diagnoses of me, my wife and my daughter were always 100% accurate [...]”*
4. **Attention:** time, questions, listens, appointment *“prompt, fast, efficient [...] he takes time for your questions [...]”*
5. **Recommendation:** years, best, family, recommend *“My mother has been a patient here for approximately 2 years [...] I have been very impressed [...]”*
6. **Competence:** know, does, not, should, help, say *“Didn’t ever really seem to know what she was talking about [...]”*

Obstetrics/Gynecology
1. **Pregnancy:** first, pregnancy, delivered, baby *“My last pregnancy was hight risk, and he stayed on top of everything [...]”*
2. **Attention:** time, questions, care, comfortable *“He spends considerable amount of time with his patients and takes extra steps [...]”*
3. **Scheduling:** office, wait, appointment, hours *“[...] she might be a good doctor, but I think she disrespects patients with the wait times.”*
4. **Manner:** caring, wonderful, great, friendly *“ The staff is great and friendly and helpful, he is the best ...”*
5. **Recommendation:** ever, years, best, love, happy *“I have been seeing her for almost seven years now and the other gyns [...] don’t even compare [...]”*
6. **Anecdotal:** surgery, results, insurance, problems *“A week later I felt sick & when I called him, he then sent me to have a sonogram [...]”*

Dentist
1. **Cost:** insurance, money, procedures, time *“First he tells you that your insurance will cover most of the cost [...]”*
2. **Manner:** staff, great, friendly, nice *“He is very considerate, nice, gentle to patients [...]”*
3. **Recommendation:** best, experience, highly *“This man is the best dentist I have ever had [...]”*
4. **Anecdotal:** teeth, crown, visit, filling *“Replaced my silver fillings with white composite ones, but had trouble [...]”*

Psychiatrist
1. **Anecdotal:** medication, told, depression, visit *“He prescribed meds for me [...] that caused me to have panic attacks [...]”*
2. **Manner:** good, caring, helpful, rude, professional *“He is truly a caring professional who always provides detailed information [...]”*
3. **Recommendation:** life, best, helped, years, feel *“I am not exaggerating be saying he has saved as well as greatly increased the quality of my life.”*
4. **Attention:** care, help, problems, work, asked *“people skills are terrible; does not listen enough, arrogant and insincere”*
5. **Schedule:** time, office, out, appointment, minutes *“You can call before you leave [...] and they will let you know how long of a wait you will have [...]”*

Figure 1. The important aspects inferred for each of the datasets. For each aspect, the label is in bold and was manually determined. The underlined words are the most frequent words determined by LDA for that aspect. The sentence in italics are extracted from the reviews and contains words associated with the aspect.

as while most reviewers have medical insurance, coverage for dental procedures is less common, and cost becomes a salient topic. Similarly scheduling is particularly salient for ObGyns and psychiatrists, but not for other specialties.

Limitations This study has a few limitations. While the use of LDA has been validated in several settings as an accurate tool for identifying topics of discussion in a large corpus of documents [10, 14, 15], in this study only a shallow manual review of the topics was carried out. In our future work, we plan to conduct a more in-depth validation of the topics with the help of a public health expert. Another limitation concerns the dataset: in our experiments, we selected reviews from a single website. Our methods can scale to a larger number of reviews and reviews from different websites. As such, this is a limitation of our experimental setup, rather than the method itself.

Future work There are many potential ways to expend this work, and we plan to investigate some of them. As rating websites become more and more popular with health consumers, there will soon be a need to provide an aggregated summary of the information provided in the many reviews available for a given provider. Assessing the quality of a review can also be carried out automatically, at least to identify coverage of aspects. If a review focuses on scheduling issues for instance, it might not be as informative as a more comprehensive review for the same provider. Finally, in addition to identifying aspects discussed in a review, methods to identify the sentiment of the reviewers for each aspect can provide additional insight about the way health consumers review their health providers and are, as such, worth investigating.

Conclusion

Content analysis of provider reviews can provide much valuable information to health consumers, health providers, health researchers and consumer health informatics researchers alike. We present a method to identify the salient aspects discussed in reviews of health providers authored by health consumers online. While there has been much work on the development and the qualitative analysis of questionnaires to assess the factors pertaining to patient satisfaction, this work takes a complimentary approach and proposes to identify the aspects that health consumers care about when choosing a health provider in a quantitative, bottom-up fashion. The aspects are learned automatically from a collection of reviews entered by health consumers, without any information other than the text of the reviews. Our findings show that such a bottom-up approach is promising, as it identifies both common and specialty-specific aspects of providers that health consumer commonly review.

Acknowledgments

This work is supported in part by a Google Research Award (NE). Any opinions, findings, or conclusions are those of the authors, and do not necessarily reflect the views of the funding organization.

References

1. W Chou, Y Hunt, E Beckjord, R Moser, and B Hesse. Social media use in the United States: Implications for health communication. *J Med Internet Res*, 11(4):e48, 2009.
2. L Frostholm, P Fink, E Oernboel, K Christensen, T Toft, F Olesen, and J Weinman. The uncertain consultation and patient satisfaction: The impact of patients' illness perceptions and a randomized controlled trial on the training of physicians' communication skills. *Psychosomatic Medicine*, 67:897–905, 2005.
3. H Rubin, B Gandek, W Rogers, M Kosinski, C McHorney, and J Ware. Patients' ratings of outpatient visits in different practice settings results from the medical outcomes study. *JAMA*, 270(7):835–840, 1993.
4. Press Ganey Associates. Medical practice pulse report: Patient perspectives on American health care. http://www.pressganey.com/galleries/default-file/2009_Med_Practice_PulseReport.pdf, 2009.
5. S O'Brien and E Peterson. Identifying high-quality hospitals: Consult the ratings or flip a coin? *Arch Intern Med*, 167(13):1342–1344, 2007.
6. B Hughes, I Joshi, and J Wareham. Health 2.0 and Medicine 2.0: Tensions and controversies in the field. *J Med Internet Res*, 10(3):e23, 2008.
7. G Eysenbach. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *J Med Internet Res*, 10(3):e22, 2008.
8. G Eysenbach. From intermediation to disintermediation and apomediation: new models for consumers to access and assess the credibility of health information in the age of Web 2.0. *Stud Health Technol Inform*, 129(Pt 1):162–166, 2007.
9. B Pang and L Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
10. D Blei, A Ng, and M Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
11. I Bhattacharya and L Getoor. A latent dirichlet model for unsupervised entity resolution. Proc. SIAM International Conference on Data Mining. 2006.
12. X Wei and W Croft. LDA-based document models for ad-hoc retrieval. Proc. of the ACM SIGIR conference. pp. 178-185. 2006.
13. L Fei-Fei and P Perona. A Bayesian hierarchical model for learning natural scene categories. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2005.
14. I Titov and R McDonald. A joint model of text and aspect ratings for sentiment summarization. Proc. of the Conference of the Association for Computational Linguistics (ACL). pp 308–316. 2008.
15. S Brody and N Elhadad. An unsupervised aspect-sentiment model for online reviews. Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp 804–812, 2010.
16. E Levine and E Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Comput.*, 13(11):2573–2593, 2001.
17. ZY Niu, DH Ji, and CL Tan. I2R: Three systems for word sense discrimination, Chinese word sense disambiguation, and English word sense disambiguation. Proc. of the International Workshop on Semantic Evaluations (SemEval), pp. 177–182, 2007.