Topic Modeling and Network Visualization to Explore Patient Experiences

Annie T. Chen, MSIS; Laura Sheble, MLIS
University of North Carolina
Chapel Hill, NC
atchen@email.unc.edu; sheble@live.unc.edu

Gabriel Eichler, PhD
PatientsLikeMe
Cambridge, MA
geichler@patientslikeme.com

Abstract—**Online support groups and health-related social networking sites can be powerful ways for patients to connect with each other and seek ways to improve their health. This content can also help clinicians and scientists understand patient experience and unmet needs. However, it can be difficult to navigate and make sense of the large volume of content. This paper proposes and demonstrates techniques that may be used to visualize topics in online health discussions. Topic modeling in conjunction with network visualization can effectively provide overviews and segmenting data. Decisions made during this process affect how content is represented and therefore, interpreted. We present a case study of a lightweight approach to model health discussion content from PatientsLikeMe using open source tools. Implications of differing representations, possibilities for extending this approach, and potential use scenarios are discussed.**

Keywords—topic modeling; network visualization; patient experience; online discussion forums; sleep

I. Introduction

Information interactions in online support forums like PatientsLikeMe (www.patientslikeme.com) assist patient learning about symptoms, treatments, and experiences of others; and may contribute to behavioral changes and outcomes [1]. Visual representations of topics and their relationships in aggregated and segmented text data can provide insight into topics discussed by patients, which offer clinicians, industry, researchers, and patients opportunities to develop new understandings and situate patient behavior in contexts through information not available from other sources such as electronic health records. We present a case study of the use of network visualizations to provide overviews of forum posts, focusing on four sleep aids: Ambien, Lunesta, trazodone and melatonin, discussed in the context of getting an adequate night's sleep.

Challenges to extracting useful information in a dataset could include the quantity of data, perceptions of data quality, data access issues, and questions related to interpretation. Our work contributes to this highly visible though sparse literature with a visually descriptive presentation of forum data, overviews, and filtering and zooming on topics of interest [2].

II. Topic Modeling and Network Visualization

Topic modeling in conjunction with network visualization is one set of techniques used to make sense of and explore large quantities of text. For example, [3] demonstrated use of topic modeling and visualization on a set of PhD thesis documents; ReportViz was developed to visualize and explore topics and keywords in public health reports [4]; and Storylines visualizes the latent semantic space of a corpus and coordinate visualizations from the perspectives of people, locations, and events involved in each story line [5].

A. Tools

Two open source packages, the MALLET toolkit [6] and Gephi [7], were utilized for topic modeling and network visualizations. To model topics, we selected a generative probabilistic modeling algorithm, Latent Dirichlet Allocation (LDA). With LDA, documents are described as random mixtures over topics; and a topic is characterized as a distribution of words [8].

B. Overviews: The Composite Network

At the outset, it may be helpful to render an overview of the content. As input, we constructed a corpus comprised of forum posts mentioning the four sleep aids: Ambien, Lunesta, trazodone and melatonin. We trained a model and predicted a set of forty topics using the MALLET toolkit. As output, we have a list of "topics" described by a set of descriptive terms; and a list of documents with estimates of their topic proportions. We manually labeled topics based on examination of term lists and documents containing high topic proportions.

We chose a two-mode network to present the relationships between posts and topics, which allowed us to discover latent topics underlying a set of discussion posts. To prepare data for rendering the network, we created a node list that includes topic and post nodes; and metadata we anticipated we would want to segment data (e.g., condition, self-reported insomnia severity level, and member post authorship experience). The edge list consists of post IDs, the topics they are connected to, and an edge weight representing the proportion of the post that contains the topic in question. The node and edge lists were imported into Gephi, to render a post-topic model (Fig. 1).

There are many ways to render and explore networks in Gephi. Here, we sized topic nodes in proportion to the number of posts to which each is connected, and the text in proportion to node size. Thus, the viewer can quickly discern the prominence of topics in relation to the broader discussion on sleep based on text size. To assist the viewer in distinguishing

This work was completed while the first author was a summer research intern at PatientsLikeMe.

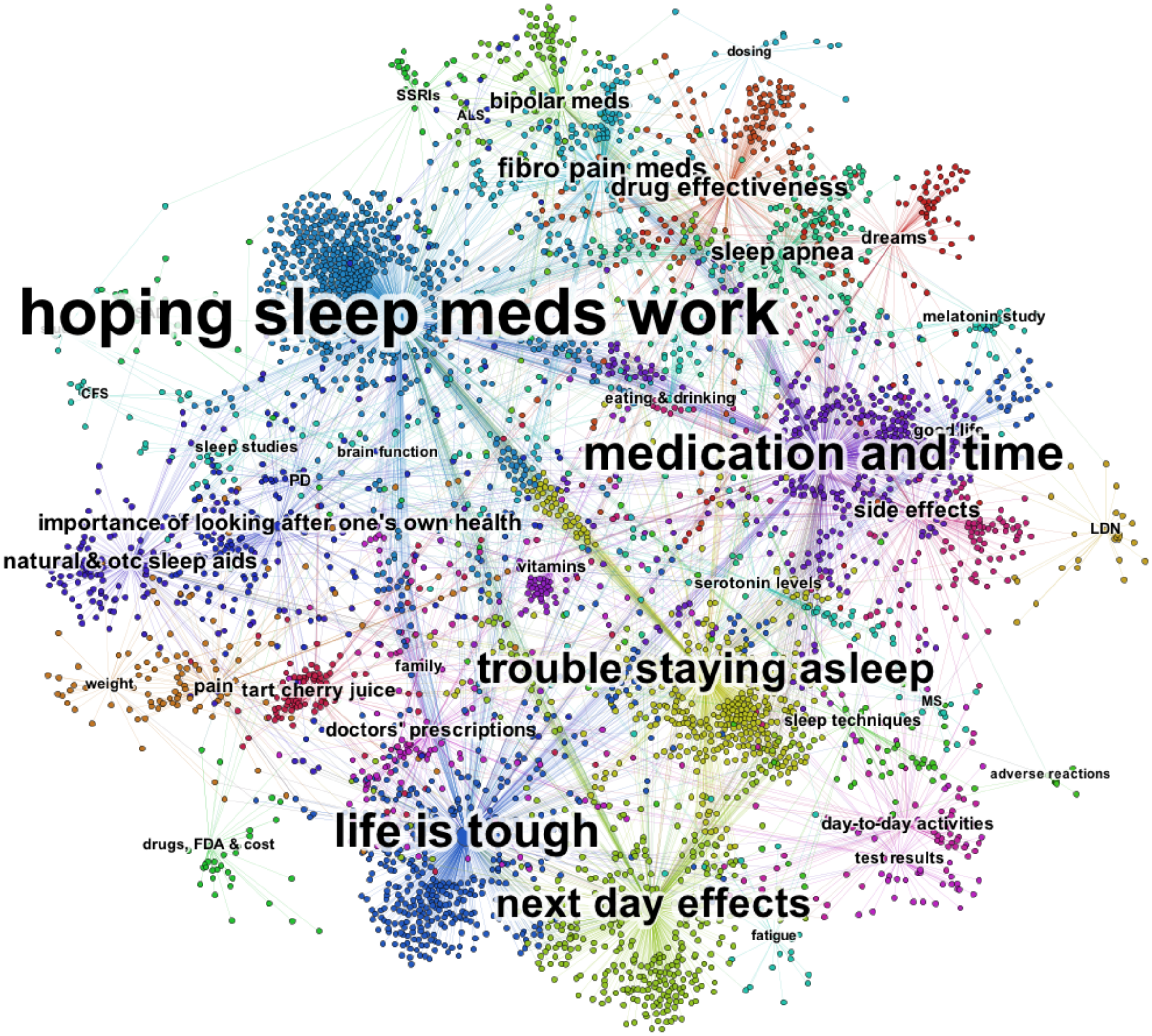


Fig. 1. Composite network of 4,405 posts mentioning four sleep aids (Ambien, Lunesta, trazodone and melatonin). Node colors reflect community membership as determined by the Louvain community detection algorithm [9].

topics, we used a community detection algorithm built in Gephi, which is based on the Louvain modularity method [9]. Node color reflects community membership.

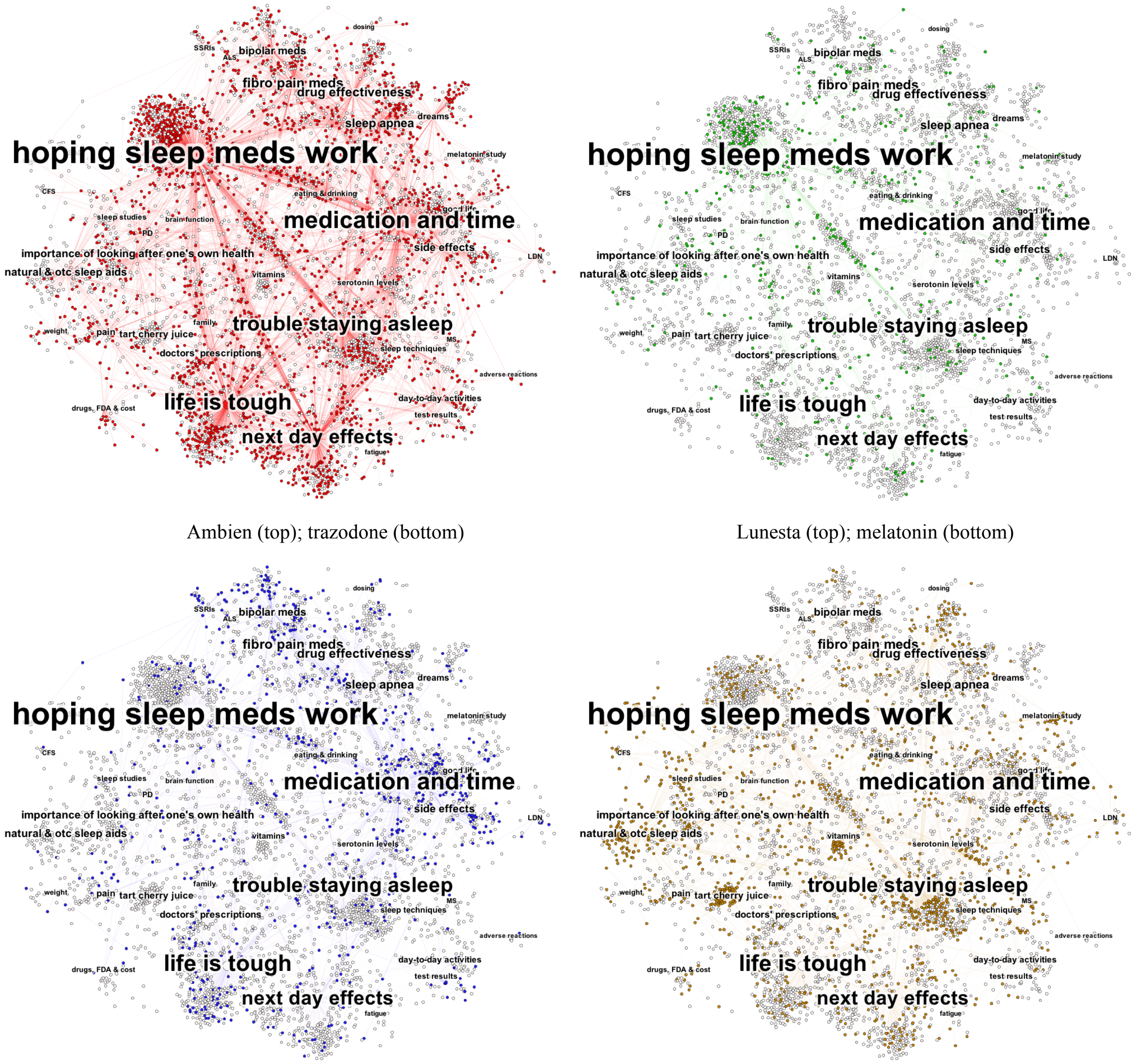
C. *Comparing Subsets of Discussion Content*

Highlighting subgraphs of a composite network can be a useful approach to research questions that focus on specific subsets of discussion posts that share a common attribute. In this case study, we were interested in exploring how conversations about these four sleep aids differed. Thus, for each sleep aid, we rendered the composite network with the subset of posts that mention the focal treatment highlighted in a distinct color, and other posts in white (at higher resolutions; but which appear gray at lower resolutions, Fig. 2).

Graphs with highlighted subsets expose differences between the highlighted and overall discussion; and when juxtaposed with other subsets, differences between conversations. In our case study, we were able to acquire a sense of the predominant topics of discussion for each sleep aid: concerns of those who talk about Ambien are similar to those of the composite network. Those who talk about Lunesta are concerned but hopeful about the effectiveness of the sleep aids. Those who discuss trazodone often mention it in the context of timing of medication use, but it also appears often in discussions about mood disorders and fibromyalgia, which makes sense given it is an antidepressant, but often taken for sleep. Melatonin is discussed in terms of its effectiveness; those who take it often find themselves still awake many hours later.

D. *Using Patient Data to Explore Associations*

Combining post content and attributes of post authors, such as their condition(s), age, and gender, and engaging in comparative evaluation may reveal new insights. In the network presented here, the proportion of patients with Multiple Sclerosis (MS) who post about "Sleep Apnea" (32%) is nearly twice the proportion of MS patients who authored posts in the network overall (18%). This finding is congruent with research that indicates a higher prevalence of sleep apnea among those with MS [10]. This retrospective evidence



Ambien (top); trazodone (bottom)

Lunesta (top); melatonin (bottom)

Fig. 2. Subsets illustrate different emphases in discussion across sleep aids.

suggests a role for social health forum data to support knowledge discovery.

E. *Applications for Forum Moderation*

Semantic network visualizations might be leveraged to aide forum moderation since structured "distant reading" can be accomplished more quickly than close readings of individual posts. Distant reading may enhance comparison across discussions, and identification of unexpected content to review at a detailed level. In our case study, we were surprised to see "Cherry Juice" was an important topic (103 posts). Almost all posts were authored by those with fibromyalgia (91%), and 67% by one person. Examining the posts, we learned about the connection – cherry juice contains melatonin, and about discussion characteristics – this sleep aid has one "champion" who writes about it often, to the point it has become eponymous (e.g., "I am going to try <champion>'s cherry juice"). Fig. 3 illustrates juxtaposing two topic clusters to highlight differences in post author attributes by topic. As mentioned, the cherry juice topic is primarily discussed by those with fibromyalgia. Trouble remaining asleep is discussed among those with a variety of conditions.

Combining topic modeling with patient condition data can help researchers understand in which subgroups information is being shared. If there is patient knowledge that may be useful to others outside the subgroup, forum moderators may choose to share this information. Examination of such patterns may also result in the discovery of novel information of interest to clinicians and/or researchers.

III. Discussion and Future Directions

Data that is organically created by patients "in the wild" can be invaluable for a wide variety of uses, including research, clinical development, patient education and forum moderation. We have shown how topic modeling and network

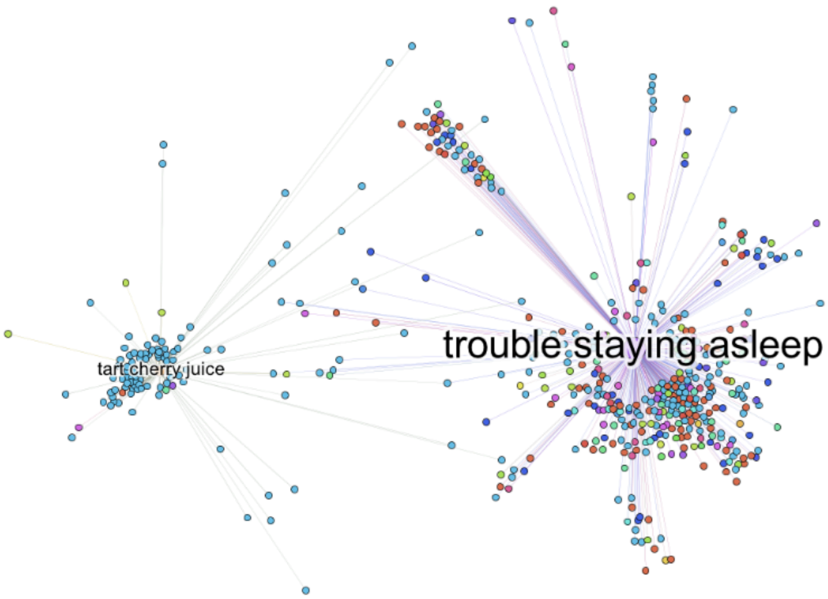


Fig.3. “Cherry juice” and “Trouble staying asleep” topic clusters. Post author conditions are identified by color (🟦=Fibromyalgia, 🟥=MS are most prevalent).

visualization might be combined to explore patient experiences. These techniques can lead to novel understandings, including about popular topics of conversation and discussion leaders. Bringing in patient attributes from other data stores can also lead to insights about the experiences of those with certain conditions. Potential uses of health discussion forum data representations are presented in Table 1.

This paper is a brief demonstration of approaches to visual exploration of forum conversations. Other directions include developing timeline visualizations to view trends in topic coverage, varying post selection criteria, aggregations by author rather than post, and connections between forum discussion topics and reports in other media sources.

TABLE I. USE SCENARIOS FOR FORUM ANALYSES

Use segments	Example scenarios
Patient education	Overview of and relationships between topics for navigation and/or education. Simplified and interactive visualizations could be presented online or integrated into PHRs
Informing clinicians	Alert services could provide information on patient concerns (e.g., side effects); and inform physicians about information patients consult online, but are unlikely to share with physicians
Clinical discovery	Discovery of novel associations (e.g., [11]) Illustrative patient experiences could be used to communicate research to broader audiences Less obtrusive data collection may enable greater participation, i.e., Citizen science [12]
Surveillance and monitoring	Overviews enable distant readings of topics associated with drugs and treatments and provide insight into patients perceptions, perhaps suggesting marketing and communication needs Contribute to diffusion analyses of health information across personal and institutional sources

REFERENCES

[1] Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, Bradley R, Heywood J. Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res 2010:12:e9.

[2] Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. Inst for Sys Res Tech Report 1996:TR 96-66.

[3] Gretarsson, B., O’Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., & Smyth, P. TopicNets: Visual analysis of large text corpora with topic modeling. J of ACM Trans on Int Sys and Tech, 2012:3: Article 23.

[4] Zhuo W. ReportViz: Interactive visualization and exploration of topics and keywords in public health reports. VAHC 2012.

[5] Zhu W, Chen C. Storylines: Visual exploration and analysis in latent semantic spaces. Comp & Graph 2007;31(3):338-49.

[6] McCallum AK. MALLET: A Machine Learning for Language Toolkit. 2002. http://mallet.cs.umass.edu.

[7] Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. Intl AAAI Conf. 2009.

[8] Blei D, Ng A, Jordan M. Latent Dirichlet Allocation. J Mach Learn Res 2003, 3:993–1022.

[9] Blondel VD, Guillaume JL, Lambiotte R, & Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theor Exp 2008;10:P1000.

[10] Braley TJ, Segal BM, Chervin RD. Sleep-disordered breathing in multiple sclerosis. Neurology 2012;*79*(9):929-36.

[11] MacLean D, Seltzer M. Mining the web for medical hypotheses. Proc HEALTHINF 2011.

[12] Swan M. Crowdsourced health research studies: An important emerging complement to clinical trials in the public health research ecosystem. J Med Internet Res 2012:14:e46.