DOI:10.1093/jncimonographs/lgt025

Understanding Topics and Sentiment in an Online Cancer Survivor Community

Kenneth Portier, Greta E. Greer, Lior Rokach, Nir Ofek, Yafei Wang, Prakhar Biyani, Mo Yu, Siddhartha Banerjee, Kang Zhao, Prasenjit Mitra, John Yen

Correspondence to: Kenneth M. Portier, PhD, Intramural Research Department, American Cancer Society Corporate Center, 250 Williams St NW, Atlanta, GA 30303 (e-mail: kenneth.portier@cancer.org).

Online cancer communities help members support one another, provide new perspectives about living with cancer, normalize experiences, and reduce isolation. The American Cancer Society's 166 000-member Cancer Survivors Network (CSN) is the largest online peer support community for cancer patients, survivors, and caregivers. Sentiment analysis and topic modeling were applied to CSN breast and colorectal cancer discussion posts from 2005 to 2010 to examine how sentiment change of thread initiators, a measure of social support, varies by discussion topic. The support provided in CSN is highest for medical, lifestyle, and treatment issues. Threads related to 1) treatments and side effects, surgery, mastectomy and reconstruction, and decision making for breast cancer, 2) lung scans, and 3) treatment drugs in colon cancer initiate with high negative sentiment and produce high average sentiment change. Using text mining tools to assess sentiment, sentiment change, and thread topics provides new insights that community managers can use to facilitate member interactions and enhance support outcomes.

J Natl Cancer Inst Monogr 2013;47:195–198

Online cancer communities provide an outlet for people with cancer and caregivers to discuss cancer-related issues. Studies of online cancer support groups and communities have shown that members benefit from online interactions in multiple ways: increased optimism (1,2), reduced stress, depression, psychological trauma (3,4), and reduced cancer concerns (5). People typically join online communities to get information and support but quickly discover that giving support to others is equally important to their survivorship (6).

The American Cancer Society's Cancer Survivors Network (CSN) (7) is the largest online peer support cancer community, with 166 000 registered members and about 25 000 unique visits a day. Although CSN supports more than 30 discussion groups, this study focuses on the two largest: the breast and colorectal cancer forums. Between 2005 and 2010, the breast and the colorectal forums had 16 604 and 12 780 threaded discussions, respectively. Discussion posts from these forums were extracted and deidentified for this study (5). Data consist of discussion threads initiated with a post from an originator to which members post replies. Threads often contain additional posts from the originator.

Sentiment analysis as used here is the automated assessment of the valence (ie, positive/negative) of posts. Sentiment analysis offers insight into the sentiments, emotions, and opinions of an online community without having to directly survey the population, a time-consuming and expensive task (8,9). Extrapolating from the buffering hypothesis (10), some of the social support provided in online communities comes from the reappraisal of a stressful event or issue (6,11), which results through community discussions and which produces a reduction in the emotional response to the event or issue. Hence, community support can be assessed using the change in sentiment between an initiating post and the first follow-up post of the initiator.

Our research goal was to examine, within the CSN community, whether sentiment change, a measure of social support, is influenced by the main topic of the initiating post (11). We hypothesized that topics that initiate with negative emotion (eg, pain, treatment, side effects) will exhibit larger sentiment change compared with topics that initiate with more positive emotion (eg, celebration), which will have smaller sentiment change. The study protocol was approved by The Pennsylvania State University institutional review board.

Previous research (12,13) used 298 CSN posts manually tagged with sentiment (*positive* or *negative*) to train a classification model that is subsequently used to assign sentiment to every post. Text features used were the following: counts of words (post length), sentences, positive sentiment words, negative sentiment words, Internet slang words, question marks, exclamation marks, and number of times a member was addressed by username or name. Derived features included average word length and ratios of the following: positive word count to post length, negative word count to post length, Internet slang word count to post length, and positive sentiment word count to negative sentiment word count. The final calibrated *AdaBoost* classification model was then used to assign sentiment for all posts and calculate sentiment change for all initiating posts (accuracy of classification: 79.2% using 10-fold cross-classification) (13). Statistical analysis found a significant positive relationship between thread originator sentiment change (average of $\text{sentiment}_{\text{self-reply}} - \text{sentiment}_{\text{initial post}}$) and the sentiment of community replies (13). This change demonstrated how originator's sentiment is positively influenced by the community.

The types of events and concerns often discussed in CSN (14) were identified using topic model analysis on thread-initiating posts. Topic classes were identified using modified latent Dirichlet allocation (LDA-VEM) (15–18). Each initiating post was assigned probabilities of belonging to each topic class. Posts were classified to the highest-probability topic. Analyses assuming 20–50 topics indicated that choice of 30 topics is reasonable for both forums. Selected word combinations identified in an initial analysis of the posts were subsequently converted to single words (eg, "breast cancer" to "breastcancer") to retain their meaning. Remaining words were reduced to root form (19), and terms occurring very often (>80% of posts) or very seldom (<5 posts) were removed before analysis.

Breast Cancer Discussions. The relationship between topics and sentiment changes is described in Figure 1, where average sentiment changes (and the associated 95% confidence interval) are plotted for each topic identified, with topics ordered from highest to lowest on their average sentiment change score. One-way analysis of variance shows significant differences among the various topic means ($F_{29,6057} = 7.39$, $P < .01$). As hypothesized, topics with lower sentiment score (greater negative emotion) for the initiating post have higher average sentiment change scores. Threads with more negative initial sentiment and higher sentiment change typically involve topics such as pain, poor laboratory results, and treatment side effects.

Colorectal Cancer Discussions. The relationship between topics and sentiment changes is described in Figure 2, which also displays differences in average sentiment change among topics overall ($F_{29,6035} = 7.39$, $P < .01$). Similar to the relationship observed in the discussion of the breast cancer forum, a negative relationship between initial sentiment and sentiment change is also evident in the colorectal cancer forum.

Both forums show that pain, medical worries, and treatment side effect issues initiate with very low sentiment and have highest sentiment change. Breast cancer posts tend to initiate with lower sentiment than colon cancer posts and sentiment change tends to be higher in breast cancer. Because the interval between the initial posts and the first follow-up responses by thread initiators is typically only 1 or 2 days, the observed sentiment change is more likely a reaction to positive sentiment posts from the community than to changes in medical status or home life.

The increased understanding of topics and related sentiment identified in this research supports the need for highly sophisticated search functionality to assist users in finding the most recent and relevant content, in addition to aiding in ongoing

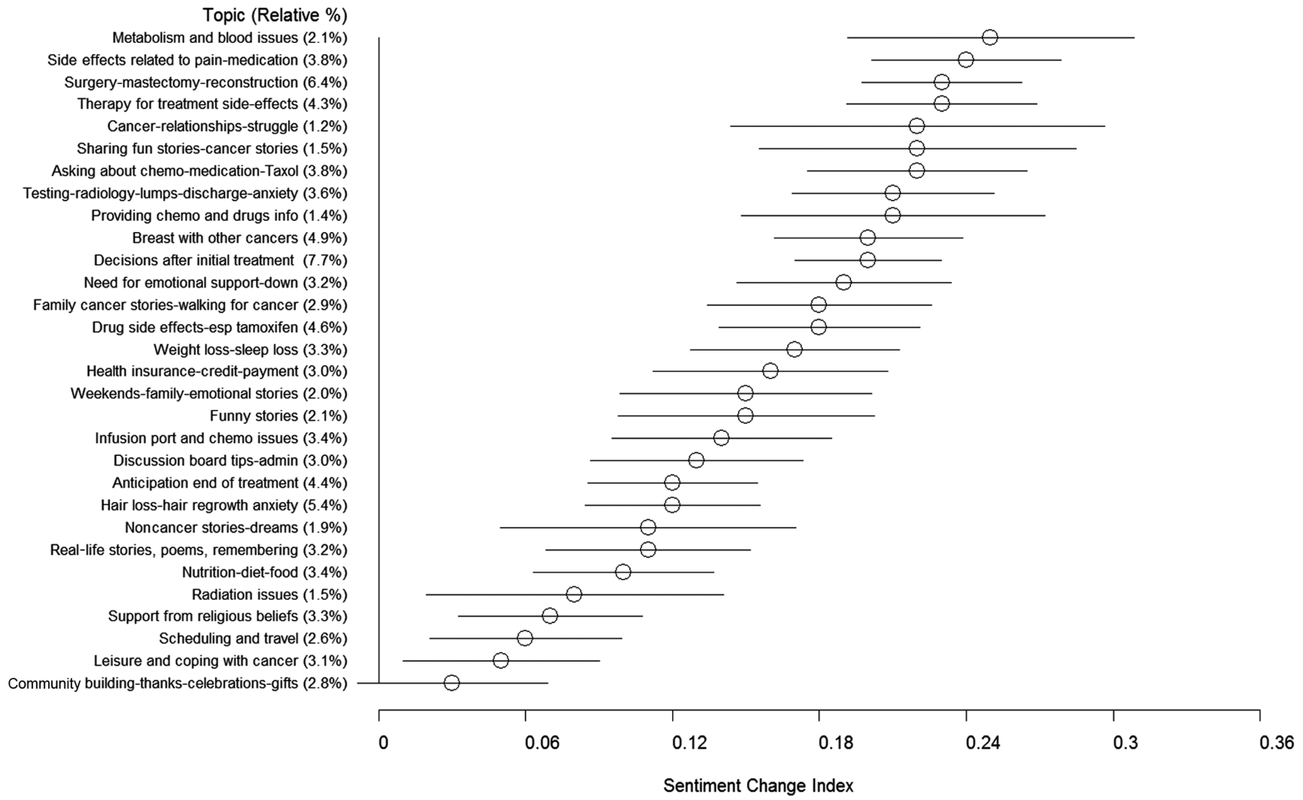

Figure 1. Average sentiment change scores with 95% confidence intervals for topics of the breast cancer discussion board of the cancer survivors network (CSN) and their relative frequencies as main post topic. High scores of average sentiment change indicate that community responses (and possibly other factors in the initiator's life) have a positive effect on the emotions of the respondent. Low scores of average sentiment change could indicate either that community response has little impact on the initiator's emotions, as represented by the sentiment of the latter's first follow-up post, or that the initial post sentiment was high to begin with, which is most likely.

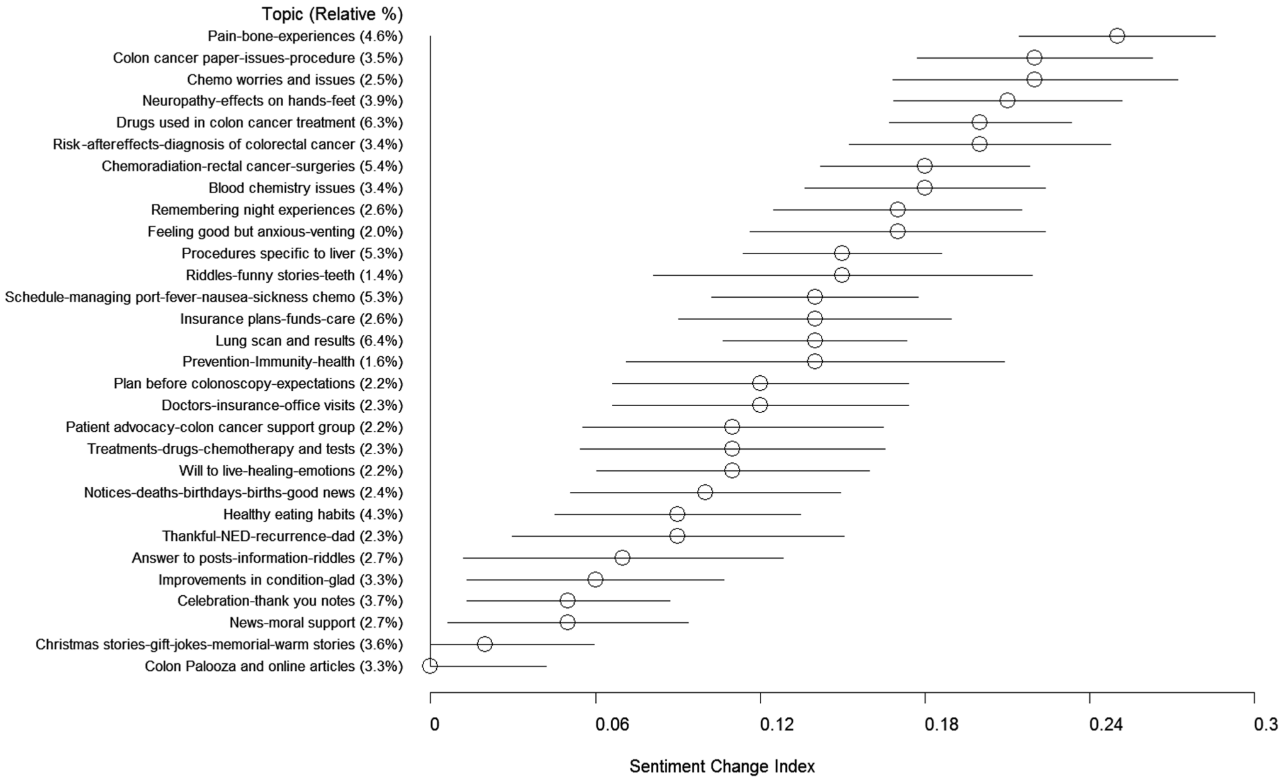

Figure 2. Average sentiment change scores with 95% confidence intervals for topics of the colorectal cancer discussion board of the cancer survivors network (CSN) and their relative frequencies as main post topic. High scores of average sentiment change indicate that community responses (and possibly other factors in the initiator's life) have a positive effect on the emotions of the respondent. Low scores of average sentiment change could indicate either that community response has little impact on the initiator's emotions, as represented by the sentiment of the latter's first follow-up post, or that the initial post sentiment was high to begin with, which is most likely. NED = no evidence of disease.

community-building efforts. Development of automated tools that monitor thread topics and associated sentiment could, for example, alert community managers to posts in need of additional community support. These types of improvements could significantly enhance social support within the community and, subsequently, members' quality of life.

References

1. Rodgers S, Chen Q. Internet community group participation: psychosocial benefits for women with breast cancer. *J Comput Mediated Commun*. 2005;10(4):00. doi:10.1111/j.1083–6101.2005.tb00268.x
2. Center for Health Enhancement Systems Studies Web site. http://chess.wisc.edu/chess. Accessed November 11, 2013.
3. Winzelberg AJ, Classen C, Alpers GW, et al. Evaluation of an internet support group for women with primary breast cancer. *Cancer*. 2003;97(5):1164–1173.
4. Beaudoin CE, Tao C-C. Modeling the impact of online cancer resources on supporters of cancer patients. *N Media Soc*. 2008;10(2):321–344.
5. Kim E, Han JY, Moon TJ, et al. The process and effect of supportive message expression and reception in online breast cancer support groups. *Psychooncology*. 2012;21(5):531–540. doi:10.1002/pon.1942.
6. Shaw BR, McTavish F, Hawkins R, Gustafson DH, Pingree S. Experiences of women with breast cancer: exchanging social support over the CHESS computer network. *J Health Commun*. 2000;5(2):135–159.
7. American Cancer Society Cancer Survivors Network Web site. http://csn.cancer.org. Accessed November 11, 2013.
8. Pang B, Lee L. Opinion mining and sentiment analysis. *Found Trends Inf Retrieval*. 2008;*2*(1–2):1–135.
9. Liu B. Sentiment analysis and subjectivity. In: Indurkhya N, Damerau FJ, eds. *Handbook of Natural Language Processing*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 2010:627–666.
10. Cohen S, Wills TA. Stress, social support, and the buffering hypothesis. *Psychol Bull*. 1985;98(2):310–357.
11. Wright K. Social support within an on-line cancer community: an assessment of emotional support, perceptions of advantages and disadvantages, and motives for using the community from a communication perspective. *J Appl Commun Res*. 2002:30(3):195–209.
12. Qiu B, Zhao K, Mitra P, et al. Get online support, feel better—sentiment analysis and dynamics in an online cancer survivor community. In: *Proceedings of The Third IEEE International Conference on Social Computing*. Boston, MA: Institute of Electrical and Electronics Engineers; 2011:274–281.
13. Zhao K, Greer G, Qiu B, Mitra P, Portier K, Yen J. Finding influential users of an online health community: a new metric based on sentiment influence. http://arxiv.org/ftp/arxiv/papers/1211/1211.6086.pdf. arXiv:1211.6086v2. Accessed November 11, 2013.
14. Rutten LJ, Arora NK, Bakos AD, Aziz N, Rowland J. Information needs and sources of information among cancer patients: a systematic review of research (1980–2003). *Patient Educ Couns*. 2005;57(3):250–261.
15. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Machine Learning Res*. 2003;3(1):993–1022.
16. McCallum AK. *MALLET: A Machine Learning for Language Toolkit* [computer program]. Amherst, MA: University of Massachusetts; 2002. http://mallet.cs.umass.edu. Accessed November 11, 2013.
17. Chang J. *Collapsed Gibbs Sampling Methods for Topic Models* [computer program]. Version 1.3.2. Vienna, Austria: The R Project; 2012. http://cran.r-project.org/web/packages/lda/lda.pdf
18. Gruen B, Hornik K. topicmodels: an R package for fitting topic models. *J Stat Software*. 2011;40(13):1–30.

19. Frakes WB. Stemming algorithms. In: Frakes WB, Baeza-Yates R, eds. *Information Retrieval: Data Structures and Algorithms*. Upper Saddle River, NJ: Prentice-Hall; 1992:132–139.

Funding

This research was supported through intramural research funding by the American Cancer Society and Pennyslvania State University.

Affiliations of authors: Intramural Research Department, American Cancer Society Corporate Center, Atlanta, GA (KP, GEG); Department of Information Systems Engineering, Ben-Gurion University of the Negev, Be'er Sheba, Israel (LR, NO); College of Information Sciences and Technology, The Pennsylvania State University, State College, University Park, PA (YW, PB, MY, SB, PM, JY); Tippie College of Business, University of Iowa, Iowa City, IA (KZ).