HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2017 November 30.

Published in final edited form as:

J Biomed Inform. 2017 May ; 69: 1–9. doi:10.1016/j.jbi.2017.03.012.

Longitudinal Analysis of Discussion Topics in an Online Breast Cancer Community using Convolutional Neural Networks

Shaodian Zhang[1], **Edouard Grave**[1], **Elizabeth Sklar**[2], and **Noémie Elhadad**[1]

[1]Columbia University, New York, NY, US

[2]King's College London, London, UK

Abstract

Identifying topics of discussions in online health communities (OHC) is critical to various applications, but can be difficult because topics of OHC content are usually heterogeneous and domain-dependent. In this paper, we provide a multi-class schema, an annotated dataset, and supervised classifiers based on convolutional neural network (CNN) and other models for the task of classifying discussion topics. We apply the CNN classifier to the most popular breast cancer online community, and carry out a longitudinal analysis to show topic distributions and topic changes throughout members' participation. Our experimental results suggest that CNN outperforms other classifiers in the task of topic classification, and that certain trajectories can be detected with respect to topic changes.

Introduction

The involvement of the Internet in healthcare gives rise to new perspectives in eHealth (Oh et al., 2005) and changes the way patients consume and contribute health-related information. Traditionally, patients with life-threatening conditions receive most of the information about their disease from their care providers. While providers tend to focus on the clinical impact of the disease and might ignore the impact of the disease on a patient's emotional wellbeing and daily life (Hartzler and Pratt, 2011), support groups, and more recently online health communities (OHCs), can act as a complementary source of support for patients (Davison et al., 2000). In particular, public online health communities such as Breast Cancer Forum (Wang et al., 2012; Elhadad et al., 2014; Zhang et al., 2014), the CSN network (Portier et al., 2013; Qiu et al., 2011), and Facebook groups (Bender et al., 2011) are getting increasingly popular among patients, and have produced unprecedented amount of user-generate content which could be valuable resources for studying OHCs.

There are many challenges in understanding the very large amount of content authored and read by online health community members, however. Some relate to the quality of information, as well as how the information is consumed and integrated by community members into their daily lives and disease management decisions. One fundamental content-related task that is important to downstream content analysis is to identify topics of discussions (Biyani et al., 2014). Previous research suggested that topic, along with

emotions, are two basic building blocks of content with respect to OHC content (Portier et al., 2013). In this study, we focus on investigating prevalences and dynamics of discussion topics in a popular online breast cancer forum. The task is challenging because topics discussed in such OHCs are usually heterogeneous and domain-dependent, and can be different from themes in other biomedical content such as clinical notes, as well as those in other types of general-purpose communities such as Facebook. Previously, topic classification has also been a central issue of text mining in general (Blei et al., 2003), but to our best knowledge no studies has been focused on automated and supervised topic modeling for online health communities.

In this paper, our study objectives are (i) to provide an annotation schema for topic classification; (ii) to contribute an annotated dataset of sentences and posts according to the coding schema; (iii) to experiment with different supervised classification tools, including convolutional neural networks, support vector machines, and labeled latent Dirichlet allocation, to automate the annotation process; and (v) to explore the prevalence and dynamics of different discussion topics in the entire breast cancer community and across member with different disease severities. Specifically, we ask following research questions:

1. What is the most effective supervised learning tool in classifying topic of discussions in an online health community?
2. What are the most prevalent topics in discussions in the breast cancer forum?
3. Are there any differences of topic foci among patients of different cancer stages?
4. How does the distribution of topics change through time, as members participate longer in the community?

1.1 Related Work

Previously, Sharf observed that in an online breast cancer group, topics regarding basic classifications or definitions of tumors and diagnosis are most prevalent, indicating that Internet support was primarily a complementary source of information in early years (Sharf, 1997). A variety of themes such as relationship/family issues became popular in online peer discussions later on (Lewallen et al., 2014; Owen et al., 2004), but disease specific topics like treatment, diagnosis, and interpretation of lab test results are still most prevalent (Civan and Pratt, 2007; Meier et al., 2007a; Cappiello et al., 2007). Specific topics of discussion were identified as well. For example, based on content analysis, Meier and colleagues found that the most common topics in 10 cancer mailing lists were about treatment information and how to communicate with healthcare providers (Meier et al., 2007a). Owen and colleagues proposed a topic schema which includes seven categories: outcome of cancer treatment, disease status and processes associated with the cancer, healthcare facilities and personnel, medical test and procedures, cancer treatment, physical symptoms and side effects, and description of cancer in the body (Owen et al., 2004). Based on such schema, prevalence of different topics can be quantified to facilitate content analysis of cancer support groups. More recently, relying on quantitative methods, topic modeling is carried out for public OHCs, but in an unsupervised fashion (Portier et al., 2013).

2 Methods

2.1 Source of data and data processing

Our work was approved by the Columbia University IRB office. We relied on the discussion board of the publicly available community from breastcancer.org. The entire content of the discussion board was collected in January 2015. The discussion board is organized in distinct forums, each with threads and posts. The following pre-processing steps were carried out.

For each post, meta-data about the forum and the thread in which it was authored was kept, along with author and creation date. The content of each post was pre-processed by (i) removing all nontextual content (e.g., substituting emoticon icons with emoticon-related codes); and (ii) identifying sentence boundaries using the open-source tool NLTK (Loper and Bird, 2002).

2.2 Creating the topic schema

To enable reliable and useful annotation of topics, we established a coding schema of discussion topics through a literature review of information needs in online health communities, with an emphasis on breast cancer communities (Meier et al., 2007b; Civan and Pratt, 2007; Blank et al., 2010; Skeels et al., 2010; Wen et al., 2011; Bender et al., 2013; Kim et al., 2013). Our objectives were (i) to devise a coding scheme that is both relevant to describing the information needs of community members as well as applicable to and robust enough for automatic topic classification; and (ii) to design a coding scheme that can be applied to characterizing topics of discussion for either an entire post or its individual sentences. Furthermore, the annotation schema is such that each unit of annotation can be labeled according to one or more topics. For instance, a given post, and even a given sentence can simultaneously convey information about a treatment and the health system.

The coding scheme was developed using an iterative process to reflect the main topics of discussion of post content. Preliminary coding of 439 sentences (corresponding to 37 posts) provided the initial categories and guidelines for coding. Upon review and discussion, infrequently used categories were collapsed into larger concepts, and the 439 sentences were coded again to verify sufficient agreement between the two initial coders. The 439 sentences and their codes were used as training instances for the later coders, along with the coding guidelines.

Our final topical scheme contains 11 topics, as listed in Table 1. It is noteworthy that the topics focus on informational support, rather than emotional dimensions and range from clinical to daily matters.

We also learned from the preliminary coding that members may shift topic of discussion in a post, which reminded us that to achieve better granularity sentence-level coding would be necessary. As such, our manual annotation described below were carried out at sentence level rather than post-level.

2.3 Manual Annotation

We selected a subset of posts (1008 posts consisting of 9016 sentences) from the original dataset described above. The posts were selected from the different forums, where each forum focuses on specific aspects of breast cancer management, such as diagnosis and treatment options, support through chemotherapy, nutrition, alternative treatments, and daily life. Posts were thus grouped in batches of 50 posts per manual annotation session.

Sentences were coded according to double annotation followed by an adjudication step from one dedicated adjudicator throughout the annotated dataset. Three coders were hired for the annotation, all female native English speakers with undergraduate degrees. To train for the annotations, coders practiced annotating the 439 sentences (37 posts) referred to above using the annotation guidelines. Inter-annotator agreement with gold-standard topic annotation was monitored throughout training, and training was terminated when a coder had achieved a 0.6 Kappa (agreement statistic) with the gold-standard annotation (Cohen and Others, 1960). Note that given the large number of potential labels in the schema and the fact that each sentence can be labeled according to multiple topics, this is a particularly stringent training constraint. Afterwards, each batch of posts was assigned two coders and was doubly annotated at the sentence level. Finally, the adjudicator went through all posts, resolved differences between coders and made final decisions over sentence topic labels.

2.4 Topic classification

Because a given sentence in a post can be described according to multiple topics (e.g., a sentence can be about a treatment, nutrition, and daily matters all at once), the task of automating the topic coding can be cast as a multi-label classification: for each sentence, there can be up to N labels, where N is the number of topics in the schema. This type of classification is more challenging than single-label classification, where one sentence can be described by only one label chosen from the N topics in a schema. Traditionally, there are two approaches for multi-label, multi-class classification: problem transformation methods and algorithm adaptation methods (Tsoumakas and Katakis, 2007).

In this paper, we rely on three different supervised classifiers, a labeled LDA classfier (Ramage et al., 2003), an SVM (Suykens and Vandewalle, 1999), and a convolutional neural network (Kim, 2014). They represent three types of mainstream supervised learning frameworks: generative graphical models, discriminative max-margin linear classifiers, and neural networks. Within these three models, labeled LDA and neural networks are able to handle multi-label classification naturally since they allow multiple outputs. For the SVM, we consider N binary, single-label classifiers and aggregates the N outputs into one multi-label.

For the labeled LDA classifier, we rely on an self-implemented Gibbs sampler for labeled LDA, based on the open source LDA implementation (Heinrich, 2005)[1]. The two hyper-parameters of the model, alpha and beta, are set as 0.1 and 0.5 experimentally. For SVM, we

[1] http://jgibblda.sourceforge.net

rely on the open source tool LibSVM (Chang and Lin, 2011), using linear kernel and all default parameters.

The convolutional neural network we used has one hidden convolutional layer. First, the sequence of words is represented as a sequence of vector of dimension $D = 100$, by using a lookup table. The word embeddings used in this lookup table were pre-trained, by using the word2vec algorithm, on the entire unannotated dataset from the same forum. Then we take the convolutions of this sequence of "word vectors" with H filters, obtaining a score for each filter and each position in the sentence. In order to obtain a fixed-size representation of the sentence, we perform max-pooling (over the positions in the sentence). We finally apply a fully connected layer to obtain a score for each topic. Since the dataset is imbalanced, we propose to use asymetric costs for positive and negative examples. The ratio between these costs is denoted by the scalar α. In our experiments, H is set to 800 and α is set to 0.25.

Prior to training the classifiers, the following pre-processing and feature selection steps were carried out: (1) all the words in the corpus were stemmed; (2) stopwords were removed from the vocabulary; (3) dimensionality reduction were carried out by doing Named Entity Recognition (using Stanford NER (Finkel et al., 2005)) to recognize Person, Location, Organization names as well as special tokens such as number, money, time. In addition, to make the comparison across tools more meaningful, we also use the word embedding input of CNN as features for SVM, examining how it differs from bag of words representations.

2.5 Application to the entire community to support longitudinal analysis

We applied the best performed classifier on all sentences in the entire unannotated dataset. For each post, we assigned it topic labels that are associated with more than 1/10 of sentences in the post. As such, based on the aggregated post-level topic labels, we are able to identify 1) what are the most prevalent topics in general in the community; 2) if there are any differences of topics among members of different cancer stages. We did not examine other factors than cancer stage in this study, because cancer stage is one particular profile information that can be accessed.

Armed with topic labels for each post in the dataset, we also conducted following longitudinal analyses to take timestamp into account. The primary objective for our analysis was to assess if participation in the community has an impact on topic of discussion. We thus compared distributions of topics of posts published in different periods of time with respect to users registration date, and tracked their changes. As such, each data point is the average frequency of a topic within all posts in a given time slice (e.g., all posts published by their authors after 3 weeks of their joining the community). To show both short-term and long-term changes, three measures of time progression are used (represented as x-axis): post, day, and week.

3 Results

3.1 Manual annotation

Table 2 shows distributions and example sentences for different topics in the manually-annotated dataset. Treatment and Miscellaneous sentences are the most frequent topics in

our annotated dataset, whereas Alternative Medicine and Test topics are the least prevalent. The high number of Miscellaneous sentences is explained by the fact that most posts start with greetings and end with encouragements, blessings, and signatures (all categorized as Miscellaneous in our coding).

Table 3 shows the inter-annotator agreement for each pair of annotators across the three annotators. Among the three coders, the first coder annotated all 1008 posts, while the other two complimentary coders are assigned part of the whole data set. The reminder of the paper reports results on the adjudicated annotation.

3.2 Topic classification

The classifiers were evaluated in a 5-fold cross validation framework using precision, recall, and F measure. In order to evaluate the overall performance of the system across all topics, micro average precision, recall and F are also calculated (Yang, 1999). Micro average takes distribution of labels into consideration, and it makes more sense in this study because of the imbalance of labels in the dataset. Experiments with a baseline system are also carried out, which simply tags every sentence with all possible labels. Aggregated results for the sentence-level classification are given in Table 4.

3.3 General prevalence of topics

Prevalence of all topics in the entire forum at post-level is given in Table 5. The most prevalent topic is personal (PERS), with 24.6% of posts labeled as such, followed by treatment (TREA, 24.6%) and diagnosis (DIAG, 9.3%). The least prevalent topics are alternative medicine (ALTR, 0.2%) and test (TEST, 1.0%). It is noteworthy that MISC did not show up in post-level annotation, because it is a default category assigned only when no other topics are identified in all sentences of the post. As such, its prevalence is extremely low at post-level and it is not of interest to our following analysis.

Clinically relevant topics such as treatment, diagnosis, and finding are more prevalent than non-clinical ones across the breast cancer forum, with one exception of PERS. Topic distribution in the entire BC dataset is more skewed that that in the annotated dataset, because the annotated dataset was sampled toward collecting more posts of rare topics such as alternative medicine (ALTR).

3.4 Topic prevalence stratified by cancer stage

In the BC dataset, many users self-reported disease information in profiles, including cancer diagnoses and treatment histories. These profile information show up in signatures when authors post, which is available to the public. In particular, out of all 57,424 authors in the dataset we crawled, 17,950 (31.3%) have their cancer stage information available in signatures. Among them, 2,325 are stage 0 (total number of posts: 170,610), 5,968 are stage I (total number of posts: 600,500), 5,907 are stage II (total number of posts: 661,990), 2,447 are stage III (total number of posts: 229,955), and 2,438 are stage IV (total number of posts: 460,313).

Topic distributions of posts published by members of different cancer stages are given in table 1. Statistical tests (multi-variate and univariate) were also carried out between numbers of different stages. Most visible differences in figure 1 are statistical significant, given relatively large sample size. Stage 0 users focus more on cancer diagnosis and health systems, which are typical topics at early times of cancer journeys. Stage IV members, counter-intuitively, discuss more about personal lives but significantly less about treatment and clinical findings. This seems to suggest that stage IV members rely on the forum to exchange emotional more than informational support with their peers. Another explanation might be that these members are so sick that few treatment options are effective for them.

3.5 Topic trajectory of users

Figure 2 shows changes of frequencies of topics after members' joining the community, in weeks, days, and individual posts, respectively. Several types of trajectories are identified. First, diagnosis is the most dominant topic at early stages of participation, especially in first posts and first days. Second, prevalence of some topics such as personal (PERS), daily matters (DAIL), and nutrition (NUTR) grow steadily, while prevalences of diagnosis (DIAG) and treatment (TREA) decline as members stay longer in the community. Third, frequencies of health systems (HSYS) and findings (FIND) increase at the beginning, but slide after reaching the peaks. Finally, alternative medicine (ALTR), laboratory test (TEST), and resources (RSRC) are unpopular topics throughout members' participation. The results seem to suggest that members' focus shifted from informational support, represented by clinically concentrated topics such as diagnosis and treatment, to emotional support, represented by personal focused on topics such as nutrition and daily lives.

4 Discussions

A wide range of topics are discussed in the breast cancer community, ranging from clinically relevant ones such as diagnosis and treatment to more daily matters such as nutritional supplements and personal lives. In the breast cancer forum, personal matters and treatment are the most dominant topics, possibly representing a mix of emotional support and informational support being exchanged.

Cancer stage plays a role in deciding members' topics of discussions. Early stage members, many of whom are newcomers to the community, care more about diagnosis related information. Stage 0 members, in particular, focus on whether certain signs indicate cancer. They also exchange anecdotes about their experiences of visiting healthcare providers when being diagnosed. Late stage members, such as stage IV members, usually stay in the community for longer time as their cancer develop. For these members, seeking information is no longer the primary motivation of participation; on the contrary, they establish closer relationships with their online peers, and disclose more personal information and support each other emotionally. It it noteworthy, however, that cancer stage information extracted from signatures may be inaccurate, since members may not report stage change timely. Also, it is naturally the case that members with late stage cancer are more likely to be long time users, which makes length of membership an important confounder in considering differences between members of different stages.

Finally, we found that members shifted their focus in participation, from clinically relevant topics to more casual topics. This coincides with the difference between cancer stages, and confirms that the difference is at least partly caused by length of participation. As members stay longer in the community and build up closer relationship with their peers, they tend to disclose more personal information, discuss more private stories, and exchange more support emotionally.

5 Conclusion

In this paper, we provide a multi-class schema, an annotated dataset, and supervised classifiers based on convolutional neural network (CNN) and other models for the task of topic classification for online health community text. We apply the classifier on the most popular breast cancer online community, the discussion boards of breastcancer.org, and carry out longitudinal analysis at scale to show topic distributions and topic changes throughout members' participation. Our experimental results suggest that CNN outperforms other classifiers in the task of topic classification. We also found that although personal and disease related topics are most prevalent, members of different cancer stages have different foci of topics. Finally, members change their interest as they participate, becoming increasingly interested in more personal topics in online discussions.

Acknowledgments

This work is supported by National Institute of General Medical Sciences Grant R01GM114355.

References

Bender, Jacqueline L., Jimenez-Marroquin, Maria-Carolina, Jadad, Alejandro R. Seeking support on facebook: a content analysis of breast cancer groups. Journal of medical Internet research. 2011 Jan. 13(1):e16. [PubMed: 21371990]

Bender, Jacqueline L., Katz, Joel, Ferris, Lorraine E., Jadad, Alejandro R. What is the role of online support from the perspective of facilitators of face-to-face support groups? A multi-method study of the use of breast cancer online communities. Patient education and counseling. 2013

Biyani, Prakhar, Caragea, Cornelia, Mitra, Prasenjit, Yen, John. Identifying Emotional and Informational Support in Online Health Communities. 2014; 1(1):827–836. anthology.aclweb.org.

Blank, Thomas O., Schmidt, Steven D., Vangsness, Stacey A., Monteiro, Anna Karina, Santagata, Paul V. Differences among breast and prostate cancer online support groups. Computers in Human Behavior. 2010; 26(6):1400–1404.

Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. the Journal of machine Learning research. 2003; 3:993–1022.

Cappiello, Michelle, Cunningham, Regina S., Knobf, M Tish, Erdos, Diane. Breast cancer survivors: information and support after treatment. Clinical nursing research. 2007 Nov; 16(4):278–93. discussion 294–301. [PubMed: 17991908]

Chang, Chih-Chung, Lin, Chih-Jen. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2:27:1–27:27.

Civan, Andrea, Pratt, Wanda. AMIA Annual Symposium Proceedings. Vol. 2007. American Medical Informatics Association; 2007. Threading together patient expertise; p. 140

Cohen, Jacob, et al. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1960; 20(1):37–46.

Davison, Kathryn P., Pen-nebaker, James W., Dickerson, Sally S. Who talks? the social psychology of illness support groups. American Psychologist. 2000; 55(2):205. [PubMed: 10717968]

Elhadad N, Zhang S, Driscoll Patricia, Brody Samuel. Characterizing the Sublanguage of Online Breast Cancer Forums for Medications, Symptoms, and Emotions. AMIA Symposium. 2014

Finkel, Jenny Rose, Grenager, Trond, Manning, Christopher. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics; 2005. Incorporating non-local information into information extraction systems by gibbs sampling; p. 363-370.

Hartzler, Andrea, Pratt, Wanda. Managing the personal side of health: how patient expertise differs from the expertise of clinicians. Journal of Medical Internet Research. 2011; 13(3)

Heinrich, G. Parameter estimation for text analysis. 2005 Sep. Web: http://www.arbylon.net/publications/textest.…

Kim, Sojung Claire, Shah, Dhavan V., Namkoong, Kang, McTavish, Fiona M., Gustafson, David H. Predictors of Online Health Information Seeking Among Women with Breast Cancer: The Role of Social Support Perception and Emotional Well-Being. Journal of Computer-Mediated Communication. 2013

Kim, Yoon. Convolutional neural networks for sentence classification. 2014 arXiv preprint arXiv: 1408.5882.

Lewallen, Andrea C., Owen, Jason E., O'Carroll Bantum, Erin, Stan-ton, Annette L. How language affects peer responsiveness in an online cancer support group: implications for treatment design and facilitation. Psycho-oncology. 2014 Jul; 23(7):766–72. [PubMed: 24519856]

Loper, Edward, Bird, Steven. Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics. Vol. 1. Association for Computational Linguistics; 2002. NLTK: The natural language toolkit; p. 63-70.

Meier, Andrea, Lyons, Elizabeth J., Frydman, Gilles, Forlenza, Michael, Rimer, Barbara K. How cancer survivors provide support on cancer-related Internet mailing lists. Journal of medical Internet research. 2007a Jan.9(2):e12. [PubMed: 17513283]

Meier, Andrea, Lyons, Elizabeth J., Frydman, Gilles, Forlenza, Michael, Rimer, Barbara K. How cancer survivors provide support on cancer-related Internet mailing lists. Journal of Medical Internet Research. 2007b; 9(2)

Oh, Hans, Rizo, Carlos, Enkin, Murray, Jadad, Alejandro. What is eHealth (3): a systematic review of published definitions. Journal of medical Internet research. 2005; 7(1)

Owen, Jason E., Klapow, Joshua C., Roth, David L., Tucker, Diane C. Use of the internet for information and support: disclosure among persons with breast and prostate cancer. Journal of behavioral medicine. 2004 Oct; 27(5):491–505. [PubMed: 15675637]

Portier, Kenneth, Greer, Greta E., Rokach, Lior, Ofek, Nir, Wang, Yafei, Biyani, Prakhar, Yu, Mo, Banerjee, Siddhartha, Zhao, Kang, Mitra, Prasenjit, Yen, John. Understanding topics and sentiment in an online cancer survivor community. Journal of the National Cancer Institute Monographs. 2013 Jan; 2013(47):195–8. [PubMed: 24395991]

Qiu, Baojun, Zhao, Kang, Mitra, Prasenjit, Wu, Dinghao, Caragea, Cornelia, Yen, John, Greer, Greta E., Portier, Kenneth. Get Online Support, Feel Better – Sentiment Analysis and Dynamics in an Online Cancer Survivor Community; 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust; 2011 Oct. p. 274-281.

Ramage, Daniel, Hall, David, Nallapati, Ramesh, Manning, Christopher D. Labeled LDA : A supervised topic model for credit attribution in multi-labeled corpora. 2003

Sharf BF. Communicating breast cancer on-line: support and empowerment on the Internet. Women & health. 1997 Jan; 26(1):65–84. [PubMed: 9311100]

Skeels, Meredith M., Unruh, Kenton T., Powell, Christopher, Pratt, Wanda. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM; 2010. Catalyzing social support for breast cancer patients; p. 173-182.

Suykens, Johan A K., Vandewalle, Joos. Least squares support vector machine classifiers. Neural processing letters. 1999; 9(3):293–300.

Tsoumakas, Grigorios, Katakis, Ioannis. Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM). 2007; 3(3):1–13.

Wang, YC., Kraut, Robert, Levine, JM. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups; Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work; 2012. p. 833-842.

Wen, Kuang-Yi, McTavish, Fiona, Kreps, Gary, Wise, Meg, Gustafson, David. From diagnosis to death: a case study of coping with breast cancer as seen through online discussion group messages. Journal of Computer-Mediated Communication. 2011 Jan; 16(2):331–361. [PubMed: 23055657]

Yang, Yiming. An evaluation of statistical approaches to text categorization. Information retrieval. 1999; 1(1-2):69–90.

Zhang S, Bantum E, Owen J, Elhadad N. Does Sustained Participation in an Online Health Community Affect Sentiment? AMIA Symposium. 2014

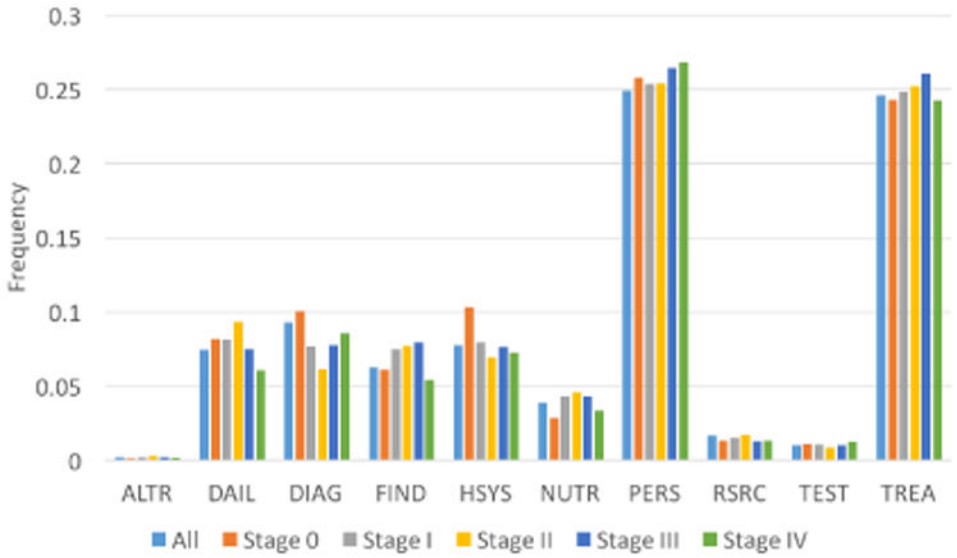

Figure 1.
Frequencies of topics of posts, stratified by cancer stages of authors.

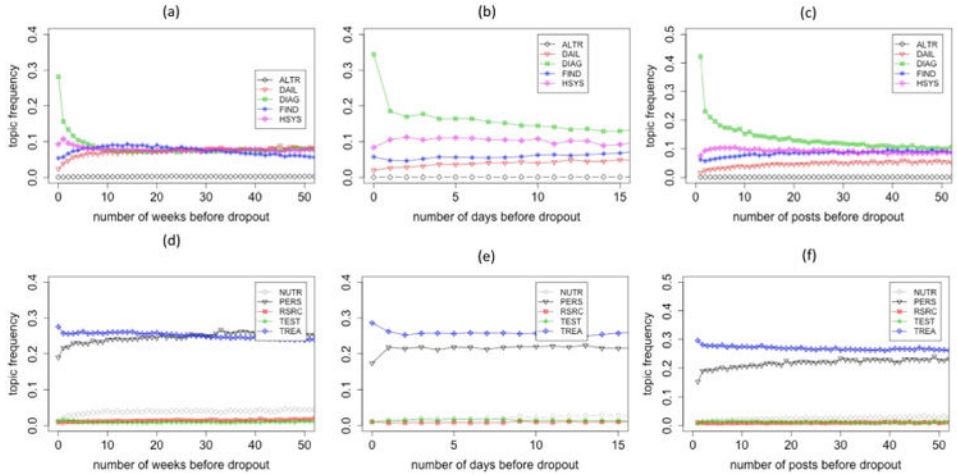

Figure 2.

How topic frequencies change through time after members join the community. X axes represents the time point after members' first activity. Y axis is the average topic frequency of all posts that are published in the corresponding time. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

Table 1
Annotation schema for breast cancer forum text

Topic	Abbreviation	Description
Alternative	ALTR	alternative and integrative medicine
Daily	DAIL	daily cancer-related experience
Diagnosis	DIAG	diagnoses, measurements, and results of tests
Finding	FIND	health finding, sign, symptom or side effect
Health Systems	HSYS	health systems patients interact with, including nurses, doctors, practices, hospitals, and insurance companies
Miscellaneous	MISC	greetings, uninformative sentence, or any sentence, which does not fit under any other annotation label
Nutrition	NUTR	nutrition
Personal	PERS	personal information
Resources	RSRC	link, pointer, or quote towards an external information resource
Test	TEST	testing procedures (but not results of tests)
Treatment	TREA	treatments, including procedures, medications and therapeutic devices

Table 2

Topic labels and the number of manually annotated sentences according to each topic. For each topic, an example of manually annotated sentence is provided. The table also includes two examples with multiple labels.

Topic	#Sentences	Example
ALTR	302	I tried everything to no avail & in desperation had acupuncture.
DAIL	600	I use virgin organic coconut oil on my skin and all organic cosmetics, shampoo, conditioner, laundry detergent, household cleaner, the works!
DIAG	1127	My cancer was a 1.2 cm mucinous bc in a duct, with low growth rate.
FIND	1195	I don't feel faint or anything- it just feels weird- anyone else out there had this happen?
HSYS	864	I don't know where you are located, but I would start with the Cancer Treatment Centers of America.
MISC	1956	Hope this helps, cheers
NUTR	608	I am staying on a bland diet, eating every 2 hours, and forcing fluids, but am worried about tomorrow based on what happened last time.
PERS	1011	He has a family history of very high triglycerides.
RSRC	568	I just did internet research and here is a good site with information on Curcumin
TEST	295	When I went in for my second mammogram on Dec. 18th, the radiologist told me I had to go get a biopsy based upon the mammogram.
TREA	2078	I'm just curious about other warriors experience with herceptin.
ALTR,NUTR	113	I read that cinnamon capsules could help with lowering glucose and ldl in our blood.
HSYS,TREA	104	After dealing with the insurance company for weeks.….she finally started taking the Xeloda last month.

Table 3

Inter-rater agreements between the three topic coders measured by Cohen's Kappa. Note that coder 1 annotated all posts while coder 2 and coder 3 annotated two complimentary parts of the data. Therefore, no agreement is calculated between coder 2 and coder 3.

Label	Coder 1 and 2	Coder 1 and 3
Avg K	0.50	0.62
ALTR	0.36	0.29
DAIL	0.30	0.50
DIAG	0.50	0.71
FIND	0.56	0.61
HSYS	0.56	0.68
MISC	0.38	0.76
NUTR	0.70	0.69
PERS	0.13	0.61
RSRC	0.63	0.58
TEST	0.69	0.70
TREA	0.67	0.71

Table 4

Topic classification performance measured by F score on different topic categories, with five classifiers. bsline: the system simply tags all sentences with all 11 labels; l-lda: the labeled LDA classifier; svm: the SVM classifier using bag of words as features; svm-e: the SVM classifier using word embedding as features; cnn: the convolutional neural network classifier

	bsline	l-lda	svm	svm-e	cnn
Micro	19.3	54.4	55.8	58.3	65.4
ALTR	6.5	9.2	9.4	30.7	35.5
DAIL	12.5	30.1	28.8	46.4	48.1
DIAG	22.2	58.8	60.2	65.3	67.1
FIND	23.4	50.1	50.9	60.0	60.3
HSYS	17.5	45.4	41.1	55.3	57.7
MISC	35.7	76.2	75.8	71.4	78.1
NUTR	12.6	57.3	58.6	68.4	72.8
PERS	20.2	24.4	26.5	47.7	47.8
RSRC	11.9	48.0	48.3	55.2	61.1
TEST	6.3	27.6	26.1	47.9	52.6
TREA	37.5	65.7	66.0	64.2	73.6

Table 5

Percentages of all topics at post level, based on automated topic classification.

ALTR	DAIL	DIAG	FIND	HSYS
0.2	7.4	9.3	6.3	7.8
NUTR	**PERS**	**RSRC**	**TEST**	**TREA**
3.9	24.9	1.7	1.0	24.6