

Comment Topic Evolution on a Cancer Institution's Facebook Page

Chunlei Tang^{1,2,3}; Li Zhou^{1,2,4}; Joseph Plasek^{1,5}; Ronen Rozenblum^{1,2}; David Bates^{1,2,3}

¹Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, USA; ²Harvard Medical School, Boston, MA, USA; ³Clinical and Quality Analysis, Partners HealthCare System, Boston, MA, USA; ⁴Clinical Informatics, Partners eCare, Partners HealthCare System, Boston, MA, USA; ⁵Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, USA

Keywords

Patient Engagement, Patient Satisfaction, Data Mining, Social Media, Consumer Participation, Oncology Service, Hospital

Summary

Objectives: Our goal was to identify and track the evolution of the topics discussed in free-text comments on a cancer institution's social media page.

Methods: We utilized the Latent Dirichlet Allocation model to extract ten topics from free-text comments on a cancer research institution's Facebook™ page between January 1, 2009, and June 30, 2014. We calculated Pearson correlation coefficients between the comment categories to demonstrate topic intensity evolution.

Results: A total of 4,335 comments were included in this study, from which ten topics were identified: greetings (17.3%), comments about the cancer institution (16.7%), blessings (10.9%), time (10.7%), treatment (9.3%), expressions of optimism (7.9%), tumor (7.5%), father figure (6.3%), and other family members & friends (8.2%), leaving 5.1% of comments unclassified. The comment distributions reveal an overall increasing trend during the study period. We discovered a strong positive correlation between greetings and other family members & friends ($r=0.88$; $p<0.001$), a positive correlation between blessings and the cancer institution ($r=0.65$; $p<0.05$), and a negative correlation between blessings and greetings ($r=-0.70$; $p<0.05$).

Conclusions: A cancer institution's social media platform can provide emotional support to patients and family members. Topic analysis may help institutions better identify and support the needs (emotional, instrumental, and social) of their community and influence their social media strategy.

Correspondence to:

Chunlei Tang, PhD
Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital
1620 Tremont Street BS-3
Boston, MA 02120, USA
Phone: (857) 600-0628
Email: ctang5@partners.org

Appl Clin Inform 2017; 8: 854–865

received: 6. April 2017

accepted in revised form: 25. June 2017

published: August 23, 2017

Citation: Tang C, Zhou L, Plasek J, Rozenblum R, Bates D. Comment Topic Evolution on a Cancer Institution's Facebook Page. Appl Clin Inform 2017; 8: 854–865
<https://doi.org/10.4338/ACI-2017-04-RA-0055>

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

1. Background and Significance

The National Cancer Institute recommends the use of patient-centered metrics to assess the quality of cancer care from a patient's perspective [1]. It has also been reported that online reviews and comments on healthcare organizations social media platforms are associated with clinical outcomes [1–12]. Monitoring social media can provide useful, unsolicited, and real-time data that might not be captured by traditional patient feedback mechanisms [2–12]. Consistent with this notion, a majority of healthcare organizations now use patient-facing health information technology tools to promote their activities and to assess their progress with respect to patient-centeredness [12]. However, recent research reveals that social media pages (such as Facebook™ and Twitter™) are primarily used by hospitals as a means to provide general information (e.g., educational) and that hospitals often don't effectively respond publicly to comments that contain patient perspectives and input [2–12].

Embracing the concept of patient engagement brings many benefits to cancer care [2–12]. Consumers are increasingly becoming more active and empowered in healthcare by sharing information on social media about personal health experiences (e.g., physicians, treatments) [2–12]. Understanding which topics are most often discussed may help health institutions better comprehend the consumer experience, engage patients and family members more efficiently, and react to the unsolicited feedback provided by consumers quicker. For example, Gage-Bouchard et al. manually examined themes for related posts (a total of 15,852 posts) through 18 publicly available Facebook pages hosted by parents of children with cancer, and identified the following themes:

1. documenting the cancer journey,
2. sharing emotional strain associated with care giving,
3. promoting awareness and advocacy about cancer,
4. fundraising,
5. mobilizing support, and
6. expressing gratitude for support [13].

Topic discovery and evolution is the task of deriving topics by either clustering documents or treating them as a multinomial distribution of words and modeling changes over time [14]. Kalyanam et al. suggests that “the most effective models developed by the topic tracking community [are] generally built on some well-known topic discovery model with a temporal aspect added to it to accommodate for the incoming stream of data [15].” Numerous full-analysis topic models have been proposed, the most popular being Latent Dirichlet Allocation (LDA) [16]. LDA is an unsupervised learning method utilizing a three-level (i.e., document, topic, word) hierarchical Bayesian model (described in detail in section 3.4) [16], and has been applied to tasks including document retrieval [17, 18], document classification [19], case-based patient retrieval [20], and therapy outcome prediction [21]. Recently, Brody and Elhadad, as well as Gao et al. analyzed national ratings data from RateMDs.com to detect salient aspects of reviews using LDA, and to study the correlation between the online rating of physicians and patients' clinical experiences, respectively [22, 23]. Wang et al. extracted pregnancy related topics using LDA and other similar algorithms from pregnancy group chat logs [24]. Yang et al. evaluated discussion threads on MedHelp and extracted topics and sentiment using LDA [25]. We perceive a gap in the patient-centered healthcare research literature surrounding automated analysis of the unsolicited but public feedback about patients' experiences on a healthcare institution social media page as opposed to healthcare provider rating sites or chat logs.

Most existing topic models lack word correlation knowledge to utilize the rich similarity relationships among words to learn coherent topics [14]. For example, given a set of comments related to the word “father”, the sense of this word can be ambiguous. Familial relationships and religion are different targets of the word “father”, which can be disambiguated by taking into account the context around the word. For example, (“children”, “daughter”) implies a familial context and (“bless”, “God”) implies a religious context. In this paper, we adapt the techniques of Xie et al. to incorporate external words to improve the coherence of LDA [14].

2. Objectives

The aims of this study were to:

1. automatically extract topics from comments posted by consumers on a cancer institutions' social media page,
2. see how the quantity of topics evolved over time, and
3. identify significant correlations between topics.

From this data, we hope to gain a more accurate understanding of oncology patients' collective care experiences. We chose a cancer institute's Facebook as our data source because it is widely used, the comments contain informative content, and the platform is timeline-based, making temporal analysis easy to conduct.

3. Methods

To investigate consumer content on a cancer research institution's Facebook page, we used a four-step, pipeline approach:

1. crawl Facebook to extract comments;
2. conduct data preprocessing to obtain eligible comments;
3. extract topics using LDA; and
4. analyze topic evolution over time.

3.1. Setting and Corpus

We collected data from the Dana-Farber Cancer Institute (DFCI)'s Facebook page. DFCI is a Boston-based cancer treatment and research center that works closely with Brigham and Women's Hospital to provide quality patient care and to conduct joint cancer research. This study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and was approved by the Partners Healthcare and DFCI Institutional Review Boards (IRB). Our IRBs perceived this particular Facebook pages comments to be provided in a public space, and therefore waived the requirement for obtaining informed consent, however we recommend those conducting similar research on online communities to consider the ethical issues related to informed consent [26].

Comments and posts published in DFCI's Facebook page between January 1, 2009, and June 30, 2014 were included in this study. Comments are composed by a Facebook user (e.g., a patient, a family member, or anyone who visits DFCI's Facebook page), and these comments may remark on a particular post published by the institution's social media engagement team. We specifically differentiate between comments and posts as these contain fundamentally different data characteristics (i.e., they are written for different purposes, posts tend to be longer in length, posts tend to contain both free-text and pictures whereas comments rarely contain pictures).

3.2. Data Crawling

Facebook comments were extracted using a self-designed web-crawler that formatted comments into a uniform layout: [Author](Comment Content, Time)]

3.3. Data Preprocessing

We conducted a data clean up step to exclude emoticons, pictures, links, and non-English text in comments. While emoticons and pictures can convey emotion, we cannot definitively define their meaning programmatically. While the other side of a link may contain useful information on which a theme can be derived, it's not something our program can handle automatically due to the added complexity of data processing (e.g., it introduces new types of data that are likely longer in length than comments on which LDA will not perform well). While non-English comments are useful to

analyze, our sample size is insufficient to include them for automatic processing. The preprocessor separated each comment by whitespace in order to get an initial word set. We filtered stop words (e.g., “an”, “the”) using Fox’s stop words list, as these are common words devoid of meaning for topic analysis and to help us to reduce the dimensionality of our data [27].

3.4. Topic Extraction

The LDA model was applied to identify the hidden topics within our Facebook comment collection [28]. We investigate only consumers’ comments (i.e., we exclude posts) for topic extraction as

1. LDA is more adept at handling the particular data characteristics of comments;
2. a comment’s topic (e.g., expressing sentiment) might deviate from their corresponding post (e.g., Institution and Staff);
3. we have a small posts dataset; and
4. our focus is on consumer feedback rather than outreach by DFCI.

LDA was chosen over other topic extraction methods because we focused on topic intensity evolution in different time slices instead of tracking changes in a single comment thread, thus we don’t need more advanced topic modeling techniques like dynamic LDA.

In LDA, each document (in this case – comment) can be presented as a multinomial distribution of topics; while each topic can be presented as a multinomial distribution of fixed word lists. LDA regards each document as a set of words (i.e., “bag of words”), thus every document can be transformed into a word frequency vector. For the purposes of our LDA model, we treat all comments as though they are independent in order to match the assumptions of LDA. Output of a LDA model is represented as topic-word probability distributions and document-topic probability distributions.

For example, suppose there are three comments A, B, and C:

- A. It needs to “spotlight” Ovarian and Thyroid cancers.
- B. My sister is a survivor.
- C. My heart and prayers go out to her son.

First, we choose a fixed number k of topics to discover. Assuming $k=2$, the topic multinomial distribution produced is 1/3 tumor and 2/3 family. Then, for each topic, the model computes the proportion of words in a document that are currently assigned to the topic (i.e., $P(\text{topic } t \mid \text{document } d)$) and the proportion of assignments to the topic over all documents that come from a word (i.e., $P(\text{word } w \mid \text{topic } t)$). Assuming additional comments within the family topic, a topic-word distribution may include the word “sister” at 30% probability, “son” at 15% probability, etc.

3.5. Hyperparameter Optimization for Topic Discovery

In LDA, determining the appropriate number of topics (k) is a fixed hyperparameter, and setting this value is more of an art than a science. K depends on the level of granularity that is meaningful for the task and the amount of information fragmentation tolerated. A tradeoff exists where increasing the number of topics leads to difficulties in interpretation and generalizability of topics discovered. For example, with a small k , we may have a single topic covering all events at DFCI, whereas if k increases, we may have separate topics for each event at DFCI (e.g., a topic for the Jimmy Fund, and another topic for the Boston Marathon), which is harder to interpret as the contents of each comment regarding an event may be similar with the only difference being the date or name, and harder to generalize to other organizations as they may not have the same types of events. Increasing the number of subtopics does not necessarily decrease the number of subtopics in the ambiguous category, as these may be smaller in quantity or salience than subtopics of existing topics (e.g., father figure as a subtopic of family). For this reason, inspired by the heuristic approach in Zhao et al. for determining the appropriate number of topics for LDA, we specified ten topics (denoted by topic k with 10 nodes labeled 0, 1, . . . , 9) [29]. Each topic contains a comment-topic probability distribution (i.e., the mixing proportions for comments), and a topic-word probability distribution (i.e., the mixing proportions for topics).

3.6 Data Analysis and Statistics

We report descriptive statistics about our corpus, including analysis of our data preprocessing results, length of comments included, and a comparison of posts quantity to comments quantity over time. We describe the topics identified by LDA, assign a theme based on the ten words with the highest probabilities for that topic, and report the quantity of posts that belong to each topic. Temporal quantity changes for topics were evaluated using the number of comments in a particular topic divided by the total number of comments included in the study. We used the Pearson correlation coefficient to find out whether different topics are correlated, where the value $r=1$ means a perfect positive correlation and the value $r=-1$ means a perfect negative correlation. We calculated Pearson correlation coefficients between the comment categories for topic intensity [30].

4. Results

From January 1, 2009, to June 30, 2014, an initial dataset of 4,849 comments from 3,318 unique authors remarking on a total of 484 posts were crawled from the Facebook site of DFCI. The data preprocessing step excluded 514 comments based on our exclusion criteria (e.g., emoticons, links, non-English text). The remaining 4,335 comments were included in this study. Length of the comments included ranged from 1 to 307 words, with a median of 9 and a mean of 11. Comments tended to be relatively short, as 80% of the comments contain less than 30 words.

In general, increases in the quantity of posts corresponded to increased commenting ($r=0.88$; $p<0.001$). The comment and post distributions reveal an overall increasing trend during the study period, but with fluctuations in a few specific time intervals (►Figure 1). For example, during the second half of year 2012 and the first half of year 2014, there were abrupt increases in comments that were likely associated with a successful social media marketing campaign associated with events happening at DFCI (e.g., Boston Marathon, fundraising initiatives like the Jimmy Fund). Prior to 2012, despite a significant investment by DFCI's staff in creating posts and responding to consumer comments, adoption of the site by other users lagged.

4.1. Topic Classification

From the ten topics discovered, we report the quantity of comments in each topic and identified a theme for each topic based on the ten words with highest probabilities that were selected by the LDA algorithm for each topic (►Table 1). A relatively specific theme for each topic was identified except topic 0 ($n=222$ (5.1%)), whose theme is ambiguous. The theme of topic 1 ($n=272$ (6.3%)) discusses emotional events concerning a "father figure" who was presumably a patient at DFCI and where the comment author is expressing gratitude and deep love for their dad. Topic 2 ($n=472$ (10.9%)) contains blessing phrases directed towards cancer patients and their families. Topic 3 ($n=326$ (7.5%)) talks about malignant tumors and a variety of cancer types (e.g., breast, lung, ovarian). Topic 4 ($n=343$ (7.9%)) conveys optimism via comment authors showing their optimistic attitudes towards beating cancer and encouraging themselves and/or other patients to fight bravely against cancer. Topic 5 ($n=356$ (8.2%)) contains stories about familial relatives of the comment author. Topic 6 ($n=405$ (9.3%)) discusses clinical treatments and includes specific mentions of treatment methods, times, places, the care team, etc. Topic 7 ($n=465$ (10.7%)) contains temporal concepts for treatment course (e.g., week, month, year). Topic 8 ($n=749$ (17.3%)) conveys holiday greetings and other date specific messages (e.g., birthdays). Topic 9 ($n=725$ (16.7%)) is related to DFCI's staff or DFCI clinics.

4.2. Topic Intensity Evolution and Interrelationships

We observed that topic 9 (DFCI) had the highest overall Pearson correlation coefficients compared to the other topics (►Figure 2), despite a drop below topics 5 (other family members & friends) and 8 (greetings) between January 2012 and June 2012.

Over the study period, we discovered a negative correlation between topic 2 (blessings) and topic 8 (greetings) ($r=-0.70$; $p<0.05$), a positive correlation between topic 2 (blessings) and topic 9 (DFCI)

($r=0.65$; $p<0.05$), and a strong positive correlation between topic 5 (other family members & friends) and topic 8 (greetings) ($r=0.88$; $p<0.001$) (►Table 2). No other relationships were significantly correlated (►Table 2).

5. Discussion

Our main finding is that most comments on the DFCI Facebook page express some form of sentiment or emotion regarding DFCI or for specific patients cared for by DFCI. Four topics (i.e., topics 1, 2, 4, 8) representing 1,835 (42.3%) comments are strongly related to human sentiment but represent different aspects of sentiment. Topics 2 (blessings) and 8 (greetings) both express blessings but are mutually exclusive, as topic 2 represents common blessings, whereas topic 8 represents date-specific blessings. The strong correlation between topics 5 (family and friends) and 8 might be the result of people who are extremely willing to convey positive-polarity sentiment for family-related experiences. Interestingly, we found an independent topic regarding a father figure, perhaps due to the higher incidence of cancer and higher mortality in males compared to women in the US, although there could be multiple other reasons for this [30]. We speculate that social media platforms (like Facebook) provide psychological comfort and encouragement for the comment authors. Patients and family members need emotional catharsis and additional support (emotional, instrumental, and social) from these platforms in order to grieve the loss of a loved one, get encouragement to continue treatment, and tend to their emotional needs. Similar to other studies, we found that social media is creating new opportunities for patients and families to share their experiences, participate actively in their care and/or in peer support groups that enable them to learn from those with similar conditions, and receive emotional support from online communities [9–10, 12].

Events at DFCI are externally hot topics among the comment authors, indicating that the DFCI's social media team has successfully engaged an interested audience with their social media outreach. Disease and treatment topics are not as readily addressed by comment authors, which may be due to a lack of posts regarding these topics to which people could react, or a lack of knowledge or interest by consumers to comment on these topics. Since social media is often seen as a valuable tool for educating the public, it is notable that it appears that most users are not using Facebook for educational purposes, although this may also have been affected by the amount of “fake news” on Facebook [30].

Recent evidence indicates that healthcare organizations and clinicians have begun to develop an interest in interacting with patients online, such as using patient surveys to collect their feedback and assessing their satisfaction [9–10, 30]. Thus, the use of social media data to detect the patient experience may supplant traditional patient surveys in the future [9–10, 30]. Our study demonstrates the value of social media content in extracting patients' experiences and provides a supplementary piece of data that acknowledges patient experiences.

5.1. Limitation

We only focused on one source of data from one particular healthcare organization. Other social media platforms may contain different types of comments and provide different insights into understanding patients' experiences. It's not clear that our results generalize beyond oncology, but it is likely that some of the non-oncology specific topics are universal. As our models don't take into account the problem of comment independence or the relationships inherent in threads of a post and related comments, we could not and do not make any claims related to the flow of these threads. Our results for Oncology Education related comments may be understated, as these types of topics are more likely to include links, which we excluded from consideration.

6. Conclusion

Our main finding is that there were more comments related to emotional interactions than engagement in discussions regarding specific disease related-problems, revealing the use Facebook as a consumer outlet for expressing emotion. Our method addresses a gap in patient-centered healthcare

research by mining the unsolicited but public feedback about patients' experiences at a healthcare institution.

Multiple Choice Questions

1. To use social media effectively, one recommended practice for healthcare organizations is to:
 - a. avoid the use of social media altogether because of privacy issues.
 - b. squelch the use of social media by employees because it can cause HIPAA violations.
 - c. only post about events at the hospital, fundraising, or educational materials.
 - d. analyze users' comments to better support patient needs.

Answer: d Explanation: Healthcare organizations should take advantage of opportunities to engage their community by using social media, analyzing and understanding what are discussed by users may help healthcare institutions better comprehend the consumer experience, engage patients and their family members more efficiently, and react to the feedback more quickly and make timely improvements to services.

Clinical Relevance Statement

The user-generated data came from online comments on a healthcare organizations social media platform, and this type of data is associated with clinical outcomes [2–12]. Emotional and psychological distress is common among loved ones of cancer patients, who sometimes report more severe mental health issues than the patients themselves. Patients' family and friends are active users of the DFCI social media page and these users tended to express a desire for support (emotional, instrumental, and social) and hope, rather than in-depth information-based content about treatments.

Conflict of Interest

The authors declare that they have no conflicts of interest in the research.

Human Subjects Protection

This study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and was reviewed by approved by the Partners and DFCI Institutional Review Boards (IRB).

Acknowledgements

We received scientific advice from several colleagues, including: Lynn A. Volk, Harpreet S. Sood, and Xiaojia Yu. The content is solely the responsibility of the authors and does not necessarily represent the official views of DFCI or Partners.

Contributions

All authors provided substantial contribution to the conception and design of this work, its data analysis and interpretation, and helped draft and revise the manuscript. All the authors are accountable for the integrity of this work.

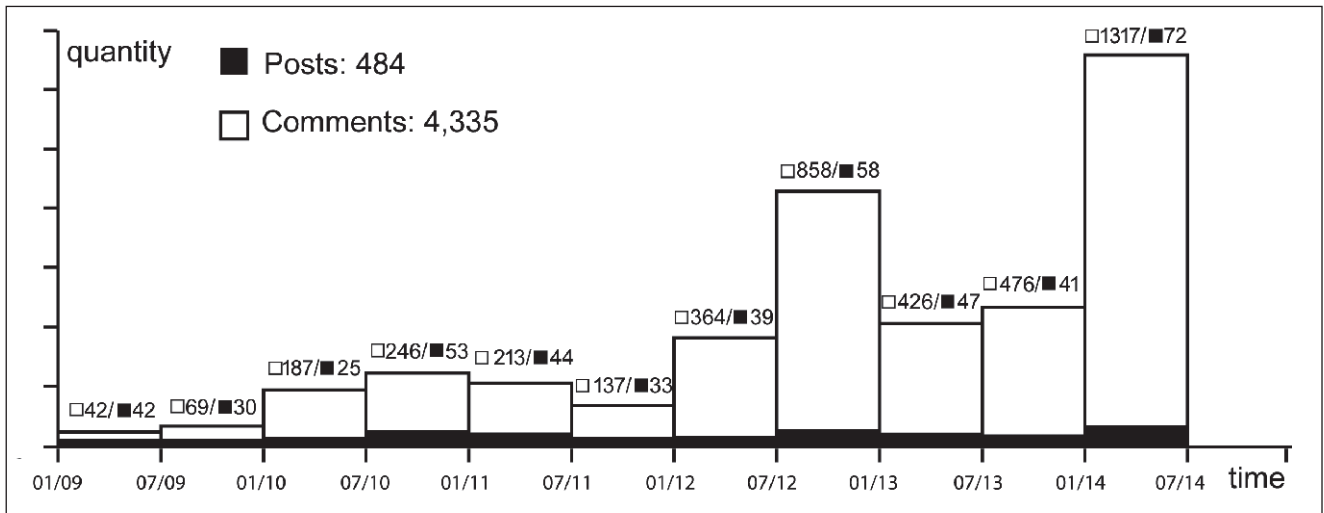


Fig. 1 Quantity comparison of posts and comments from DFCI. Strong positive correlation exists between comment and post quantity (Pearson correlation coefficient: $r=0.88$; $p<0.001$)

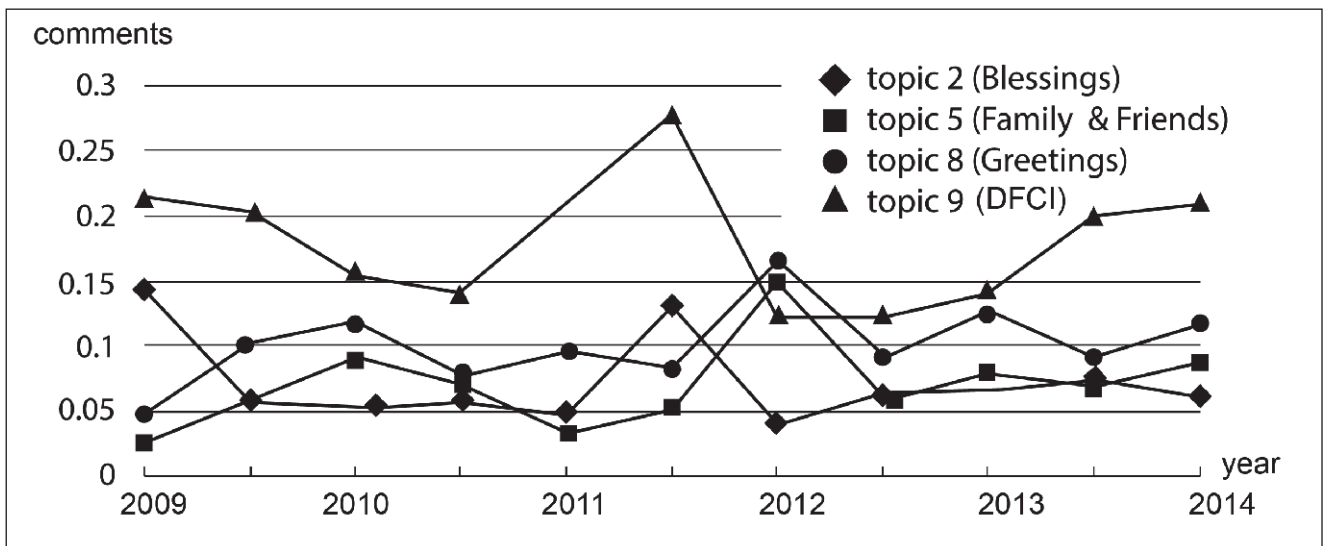


Fig. 2 Topic intensity evolution. Lines represent Pearson correlation, not frequency.

Table 1 Topic Classification (sorted by number of comments); ^aA total of 4,335 comments were included in this study. ^bComments have been modified slightly to preserve anonymity. ^cConveys human emotion.

Topic	n ^a (%)	Theme	Top 10 Words	Representative Comment Examples ^b
8 ^c	749 (17.3)	Greetings	Great / day / happy / today / birthday / kids / congratulations / wonderful / thankful / hugs	... Have a great day!HAPPY EARTH DAY!Congratulations on today's successful marathon completion...
9	725 (16.7)	Institution and staff	Dana / Farber / care / staff / patient / doctor / hospital / nurse / research	...Is Dana Farber Cancer Institute aware of this The focus on patient care has always been wonderful and energizing From doctors, nurses, and every staff are well professionals...
2 ^b	472 (10.9)	Blessings	God / bless / life / beautiful / prayers / family / angel / rest / heart / peace	...God bless her family and her... ...R.I.P... ...My heart and prayers go out to her family...
7	465 (10.7)	Time	Years / chemo / long / day / month/ week /ago / cure / yrs / forever	...They added another 20 years to my life... ...to donate once a month for 6 day then week off and so on for a long period of time....
6	405 (9.3)	Treatment	Time / treatment / amazing / people / place/ team / work / year / transplant / blood	...before, during, and after treatment.... ...to Yawkey, the best needle team in the universe,had a bone marrow transplant...
5	356 (8.2)	Other family members and friends	Family / support / friends / mom / son/ children / sister / fund / brother	...I've lost some dear friends to cancer, two Aunts, both Grandmothers and my maternal Grandfather. My mom, two first cousins and several friends have been diagnosed with cancer... ...My sister and I both got tattoos...We didn't tell dad or any members of our family... ...My daughter is the survivor...
4 ^c	343 (7.9)	Optimism	Fight / live / strong / stay / positive/ thing / healthy / pray / young / faith	...she was so strong and such a fighter.... Stay Positive...Positive...Positive!the more positive out come your going to get...keep your head up but there isn't any better feeling than some encouragement or motivation to fight it from a loved one or a close friendI will fight this and live to see another day...
3	326 (7.5)	Tumor	Cancer / breast / lung / survivor / free/ stage / ovarian / die/ diagnosis / passed	...have acknowledged that the birth control pill has a cancer-causing carcinogen... ...being overweight or obese, which is a risk factor for breast cancer... ...need to "spotlight" Ovarian and Thyroid cancers...
1 ^c	272 (6.3)	Father figure	Love / father/ awesome / nice / sweet/ make / dad / nice / guys / person	...He was an awesome father and an admirable man... ...I love my dad! He is the best father I could have ever had..... ...the best lesson a dad can teach their daughter is...
0	222 (5.1)	Ambiguous	Good / hope / give / people / feel/ back/ luck/ work / home / lot	...good docs know this is true... ...with kindness, respect and has continued to give us hope!I hope that it will start people conversing about cancer....

Table 2 Interrelationships between Topics

Topic A	Topic B	Pearson r	p value
topic 2 (blessings)	topic 5 (family and friends)	-0.58	0.06
topic 2 (blessings)	topic 8 (greetings)	-0.70	<0.05
topic 2 (blessings)	topic 9 (DFCI)	0.65	<0.05
topic 5 (family and friends)	topic 8 (greetings)	0.88	<0.001
topic 5 (family and friends)	topic 9 (DFCI)	-0.56	0.07
topic 8 (greetings)	topic 9 (DFCI)	-0.50	0.11

References

1. Assessment of Patients' Experience of Cancer Care (APECC) Study. National Cancer Institute; 2017 [cited 3/27/2017 3/27/2017]; Available from: <https://healthcaredelivery.cancer.gov/apec/>.
2. Greaves F, Pape UJ, King D, Darzi A, Majeed A, Wachter RM, Millett C. Associations between Web-based patient ratings and objective measures of hospital quality. *Arch Intern Med* 2012; 172(5): 435–6.
3. Bardach NS, Asteria-Penaloza R, Boscardin WJ, Dudley RA. The relationship between commercial website ratings and traditional hospital performance measures in the USA. *BMJ Qual Saf* 2013; 22(3): 194–202.
4. Greaves F, Millett C. Consistently Increasing Numbers of Online Ratings of Healthcare in England. *J Med Internet Res* 2012; 14(3): e94.
5. Munson AS, Cavusoglu H, Frisch L, Fels S. Sociotechnical Challenges and Progress in Using Social Media for Health. *J Med Internet Res* 2013; 15(10): e226.
6. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. *J Med Internet Res* 2013; 15(11): e239.
7. King D, Ramirez-Cano D, Greaves F, Vlaev I, Beales S, Darzi A. Twitter and the health reforms in the English National Health Service. *Health policy (Amsterdam, Netherlands)* 2013; 110(2–3): 291–7.
8. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Quality & Safety* 2013; 22(3): 251–5.
9. Rozenblum R, Greaves F, Bates DW. The role of social media around patient experience and engagement. *BMJ Qual Saf* 2017 Apr 20.
10. Rozenblum R, Bates DW. Patient-centred healthcare, social media and the internet: the perfect storm? *BMJ Qual Saf* 2013; 22: 183–6.
11. Hawkins JB, Brownstein JS, Tuli G, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf* 2016; 25: 404–13.
12. Rozenblum R, Miller P, Pearson D, Marielli A. Patient-centered healthcare, patient engagement and health information technology: the perfect storm. In: Grando M, Rozenblum R, Bates DW, eds. *Information Technology for Patient Empowerment in Healthcare*. 1st ed. Berlin: Walter de Gruyter Inc; 2015: 3–22.
13. Gage-Bouchard EA, LaValley S, Mollica M, Beaupin LK. Cancer Communication on Social Media: Examining How Cancer Caregivers Use Facebook for Cancer-Related Communication. *Cancer nursing* 2017; Publish Ahead of Print.
14. Andre Gohr AH, Rene Schult , Myra Spiliopoulou. Topic evolution in a stream of documents. In *SDM* 2009; 859–72.
15. Kalyanam J, Mantrach A, Saez-Trumper D, Vahabi H, Lanckriet G. Leveraging Social Context for Modeling Topic Evolution. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Sydney, NSW, Australia. 2783319: ACM 2015; 517–26.
16. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003; 3: 993–1022.
17. Liang S, Yilmaz E, Kanoulas E. Dynamic Clustering of Streaming Short Documents. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; San Francisco, California, USA. 2939748: ACM 2016; 995–1004.
18. Amoualian H, Clausel M, Gaussier É, Amini M-R, editors. *Streaming-LDA: A Copula-based Approach to Modeling Topic Dependencies in Document Streams*. KDD 2016.
19. Sarioglu E, Yadav K, Choi H-A, editors. *Topic Modeling Based Classification of Clinical Reports*. ACL (Student Research Workshop) 2013.
20. Arnold CW, El-Saden SM, Bui AA, Taira R, editors. *Clinical case-based retrieval using latent topic analysis*. AMIA Annual Symposium Proceedings; 2010: American Medical Informatics Association.
21. Howes C, Purver M, McCabe R, editors. *Investigating Topic Modelling for Therapy Dialogue Analysis*. Proceedings IWCS Workshop on Computational Semantics in Clinical Text (CSCT) 2013.
22. Brody S, Elhadad N. Detecting Salient Aspects in Online Reviews of Health Providers. *AMIA Annual Symposium Proceedings* 2010; 2010: 202–6.
23. Gao GG, McCullough SJ, Agarwal R, Jha KA. A Changing Landscape of Physician Quality Reporting: Analysis of Patients' Online Ratings of Their Physicians Over a 5-Year Period. *J Med Internet Res* 2012; 14(1): e38.
24. Wang T, Huang Z, Gan C. On mining latent topics from healthcare chat logs. *J Biomed Inform* 2016; 61: 247–59.
25. Yang FC, Lee AJ, Kuo SC. Mining Health Social Media with Sentiment Analysis. *J Med Syst* 2016; 40(11): 236.
26. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. *BMJ* 2001; 323(7321): 1103–5.

27. Fox C. A stop list for general text. SIGIR Forum 1989; 24(1-2): 19-21.
28. Mei Q, Zhai C. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining; Chicago, Illinois, USA. 1081895: ACM 2005; 198-207.
29. Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, Zou W. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics 2015; 16(Suppl 13): S8-S.
30. Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. Noise reduction in speech processing: Springer 2009; 1-4.