Get Online Support, Feel Better—Sentiment Analysis and Dynamics in an Online Cancer Survivor Community

Baojun Qiu*, Kang Zhao†, Prasenjit Mitra†, Dinghao Wu†, Cornelia Caragea†, John Yen†,
Greta E Greer‡, Kenneth Portier‡
*Department of Computer Science and Engineering
†College of Information Sciences and Technology
The Pennsylvania State University, University Park, PA 16802
‡American Cancer Society, Inc.
250 Williams Street NW, Atlanta, GA 30303
Email: {bqiu, kxz134, pmitra, dwu, ccaragea, jyen}@ist.psu.edu, {greta.greer, kenneth.portier}@cancer.org

***Abstract*—Many users join online health communities (OHC) to obtain information and seek social support. Understanding the emotional impacts of participation on patients and their informal caregivers is important for OHC managers. Ethnographical observations, interviews, and questionnaires have reported benefits from online health communities, but these approaches are too costly to adopt for large-scale analyses of emotional impacts. A computational approach using machine learning and text mining techniques is demonstrated using data from the American Cancer Society Cancer Survivors Network (CSN), an online forum of nearly a half million posts. This approach automatically estimates the sentiment of forum posts, discovers sentiment change patterns in CSN members, and allows investigation of factors that affect the sentiment change. This first study of sentiment benefits and dynamics in a large-scale health-related electronic community finds that an estimated 75%–85% of CSN forum participants change their sentiment in a positive direction through online interactions with other community members. Two new features, *Name* and *Slang*, not previously used in sentiment analysis, facilitate identifying positive sentiment in posts. This work establishes foundational concepts for further studies of sentiment impact of OHC participation and provides insight useful for the design of new OHC's or enhancement of existing OHCs in providing better emotional support to their members.**

I. Introduction

Cancer accounted for nearly one in every four deaths in the United States [1], and approximately 13% or 7.4 million deaths worldwide in 2007. According to the World Health Organization [2], cancer is estimated to cause 12 million deaths worldwide in 2030. Today, about 12 million Americans either have recently diagnosed with cancer or identify themselves as a cancer survivor [1]. Many survivors and their family or friend caregivers experience not only physical effects but also emotional effects such as stress, anxiety, and depression [3] from cancer and cancer treatments.

Each year, many people turn to the Internet to satisfy their health-related needs [4] for information and support. A 2010 study by the Pew Research Center found that 83% of American adult Internet users utilize the Internet for health-related purposes [5], with more than 25% of these users [6] seeking social support through joining and participating in an online health community (OHC). Unlike websites that only offer slowly changing medical or other health-related information, an OHC typically includes features such as discussion boards, chat rooms, etc. where users can interact with each other. Support and information from people with similar cancers or problems is very valuable because cancer experiences are unique, and family members, friends and cancer care providers often do not understand the problems [7]. A cancer OHC that includes both survivors and their caregivers offers a way to share experiences about their cancer and cancer treatment, seek solutions to daily living issues, and in general, support one another [8] in ways that are not often possible with other close family, friends or even health care providers.

Benefits to cancer survivors who have participated in an OHC are reported in the literature. OHC participation increases social support [9][10], reduces levels of stress, depression, and psychological trauma [11][12], and helps participants be more optimistic about the course of their life with cancer [10]. The support received from other OHC members help cancer patients better cope with their disease and improve their lives both physically and mentally [9][13]. Caregivers for cancer patients typically receive similar benefits.

Most previous research is based on data collected and analyzed using traditional social science methods, such as ethnographical observations, interviews, questionnaires, surveys, and statistical testing. These data collection methods pose three challenges. First, the scale of the data is limited due to the fact that direct observation and interview takes a lot of time. Obtaining data from the thousands of users in even a moderately-sized OHC is expensive, resource-intensive and impractical. Second, the sample used is typically biased. For example, active members who are happy and satisfied with an OHC are more likely to respond to researchers' questionnaires or surveys than those who are inactive or unhappy with the community. In addition, much of the data are dependent on personal recall of past events. A person's recall of past events

and emotions is often flawed or incomplete. Remembering emotional reactions to other OHC member responses to one or more topics or questions posted sometime in the past is unlikely to be accurate [14][15]. Third, these methods often have coarse temporal granularity. Tracking real-time emotional dynamics in direct association with OHC participation is extremely difficult with these methods.

In this research, a computational social science [16] approach to the study of how cancer survivors and caregivers benefit from participations in an OHC is presented. Computing technologies enable recording and analysis of the asynchronous and distributed social interactions in an OHC, making large amounts of data available for analysis. Computation also makes it possible to analyze the content and structure of online interactions captured in these large-scale data. Using a popular OHC for cancer survivors as a case study, discussion forum interactions for a 10-year time period are analyzed to provide insight into how cancer survivors' and caregivers' participation in the OHC affect emotions on a larger scale and in a systematic fashion. In what follows, the OHC forum used is described and basic use statistics provided. Next, machine learning techniques [17] to analyze sentiment of forum discussions are illustrated. A new approach to analysis of sentiment dynamics and identification of contributing factors are presented followed by conclusions and discussion of future research needs.

II. Case Study Data: American Cancer Society Cancer Survivors Network

The American Cancer Society, a national community-based volunteer health organization, designed and maintains its Cancer Survivors Network (CSN) (http://csn.cancer.org) as a dynamic online community for cancer patients, cancer survivors and their families and friends. It is considered a safe and welcoming place for them to support one another and share their cancer or caregiving experiences, feelings and practical tips for dealing with many of the issues encountered during cancer treatment and subsequent survival. Launched in July of 2000, it currently has more than 137,000 member participants.

The conceptual framework for CSN is based on the work of Irvin D. Yalom, M.D., a respected group work theoretician and practitioner [18]. By design, CSN provides peer support and psychosocial intervention services, utilizing group dynamics to facilitate therapeutic factors such as the instillation of hope, universality, catharsis, existentialism, altruism, interpersonal learning and group cohesiveness. The most commonly used CSN feature for group interaction is the *forum* consisting of 38 discussion boards of which 25 are cancer-specific. The breast cancer and colorectal cancer boards are the two most active discussion boards. Non-cancer-specific boards are devoted to topics such as humor, caregiving, emotional support, and spirituality.

Forum posts from July 2000 to October 2010 comprising 48,779 threads and more than 468,000 posts from 27,173 participants were downloaded into a dataset. Participants-respondents are identified by code only and this approach passed IRB review. A nationwide ACS marketing campaign featuring CSN conducted in the spring of 2008, resulted in a greater than 300% growth in CSN membership with the result that about 70% of all forum posts occur after this campaign. Forum post rates are relatively stable from May 2009 to October 2010 at an average of 16,550 posts per month. Participant total post counts, including initial posts and all replies, follows a power-law distributions (see Figure 1), demonstrating what is typical for most on-line communities, that most participants publish few posts while a few highly active participants publish a large number of posts.

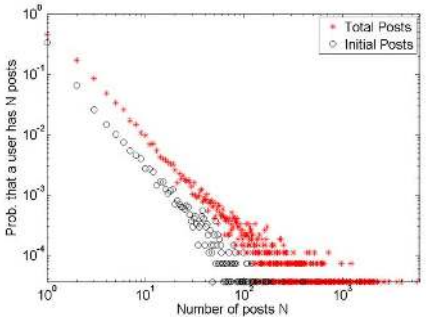

Fig. 1. Distributions of CSN participants' contribution (initial posts and total posts) to the online forum.

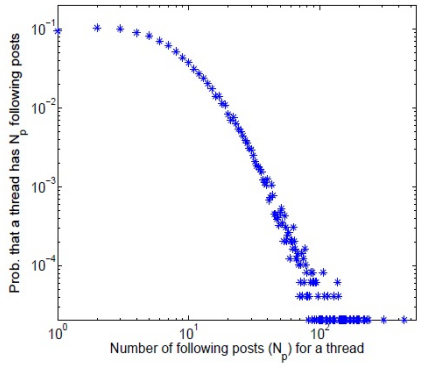

Fig. 2. Distribution of the number of replies to threads in CSN forums.

Forum posts are organized in threaded discussions (threads) comprised of initial posts and replies. The number of replies in a thread also approximately follows the power-law distribution (see Figure 2). The life span of a thread, defined as the time between the initial post and last reply is distributed as shown in Figure 3. This right skewed distribution estimates mean thread life span of 1,725 hours (or about 72 days) but median of only 58 hours (2.4 days) with 72% of threads having life spans shorter than seven days. The correlation between the number of replies and the life span of a thread is weak with Pearson correlation coefficient of 0.04. This suggests that long threads having many replies do not necessarily have long life span. Basic statistics of the forum data used in this analysis are presented in Table I.

III. Sentiment Analysis

The sentiment of a participant's post, reflecting their emotion at the time of posting, is not directly observable, and

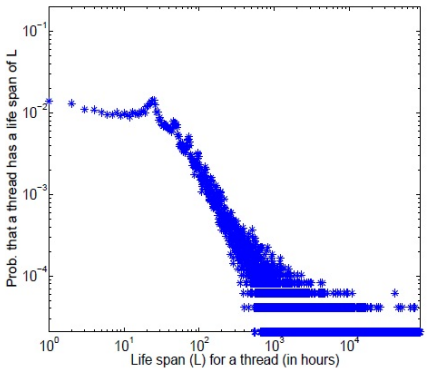

Fig. 3. Distribution of the time span of threads in CSN forums.

TABLE I
SUMMARY STATISTICS FOR CSN FORUM THREAD DATA.

	Mean	Median	Maximum
Number of posts by a participant	17.25	2	5,607
Number of replies per thread	8.7	6	442
Life span of a thread	1,725 hours (71.9 days)	58 hours (2.4 days)	87,846 hours (10 years)

hence must be modeled as a latent variable. The sentiment of a participant who initiates a thread (the originator) can be analyzed at both the beginning and the end of a thread. In this way, it is possible to determine whether the support provided by participants who responded to the thread (the respondents) is able to change the sentiment of the thread originator. Manually labeling sentiment levels for tens of thousands of posts is not feasible. Automatic identification methods previously developed for sentiment analysis [19][17] are used instead. These methods use machine learning and text mining [20] to identify individual opinions in text and use them to illustrate and analyze the dynamics of the underlying sentiment. Use of this methodological approach to sentiment analysis is growing, with recent applications made to recommender systems, business and government intelligence, and computational politics [19][17].

Most computational sentiment analysis is focused on selecting indicative lexical features that allow classification of texts into *positive* or *negative* [17] sentiment classes. This task is made easy when sentiment-labeled data is readily available, such as with product marketing data[17] where consumers are directly asked about positive (desirable) or negative (undesirable) aspects of a product. However, what is considered desirable in one domain may not be so in another domain. Consider for example the use of the word "positive". In the cancer domain, one can have a "positive" diagnostic test which might indicate the presence of cancer (an undesirable sentiment) or might indicate that the cancer is responding to therapy (a desirable sentiment). This implies that training data must, by design, be unique to the domain being analyzed and take into account context when classifying text sentiment. At the same time, no one classification approach will work in all

TABLE II
EXAMPLES OF CSN POSTS AND THEIR CLASSIFICATION.

Label	Post
Negative	My mom became resistant to carbo after 7 treatments and now the trial drug is no longer working :(, ...
Positive	ID-x, I love the way you think, ..., hope is crucial and no one can deny that a cure may be right around the corner!!!

situations. The classification model must also be adapted to take into account text structure of the domain.

Training data for the CSN breast cancer and colorectal cancer forums were created by first manually labeling several hundred posts in each forum into *positive* or *negative* sentiment classes. Next, features were extracted from these posts. Finally, classification models were trained (fit) to these data. The goal is to produce a fitted classification model that perfectly classifies posts back to their original defined *positive* or *negative* sentiment class using only the extracted features. The fitted sentiment classification model is then applied to all unlabeled posts allowing each to be classified to a sentiment class. These classified data are then used to investigate sentiment dynamics in the CSN forum.

A. Text Mining and Feature Extraction

A simple random sample of 298 posts was selected from the CSN breast cancer forum and each post manually classified as being of *positive* or *negative* sentiment with the result that 204 of them were labeled as *positive* and 94 were *negative*. An example of a negative and of a positive post is shown in Table II. In the positive example, ID_x is the identifying code for a unique CSN participant.

Next, lexical and style features were extracted from the posts using standard text mining techniques. Features defined and extracted from the data are summarized in Table III. *Pos* and *Neg* labels contain the numbers of positive and negative words (and emoticons) respectively in a post. The positive and negative word lists used to produce these counts are from Hu and Liu [21], and the positive and negative emoticon lists were collected from the Internet (http://en.wikipedia.org/wiki/List_of_emoticons). Many posts in the CSN forum mention names, e.g., *ID_x, I love the way you think*. To facilitate assessment of whether name mention has a relationship with sentiment, the feature *Name* contains a count of the occurrences of coded names in the post. The feature *Slang* contains the number of slang words used in the post. Thelwall *et al.* [22] introduced two additional features, *PosStrength* and *NegStrength* which were also measured for each post. Different from *NumOfPos* and *NumOfNeg*, *PosStrength* and *NegStrength* considers, in addition to the fact that a word is in the positive or negative word lists, the strength of emotion displayed. For example, *very good* and *good!!!* are scored as more positive than *good*. Learning algorithms [22] are used to establish the strength of individual words. The lists utilized at this stage of analysis are available at http://sites.google.com/site/qiubaojun/psu-sentiment.zip.

TABLE III
FEATURES FOR A POST

Feature	Definition
PostLength	The number of words
Pos	NumOfPos/PostLength, where NumOfPos is the number of positive words/emoticons
Neg	NumOfNeg/PostLength, where NumOfNeg is the number of negative words/emoticons
Name	NumOfName/PostLength, where NumOfName is the number of names mentioned
Slang	NumOfSlang/PostLength, where NumOfSlang is the number of Internet slangs
PosStrength	Positive sentiment strength [22]
NegStrength	Negative sentiment strength [22] (The value is more below 0 if the sentiment strength is more negative)
PosVsNeg	(NumOfPos+1)/(NumOfNeg+1)
PosVsNegStrength	PosStrength/NegStrength
Sentence	The number of sentences
AvgWordLen	The average length of words
QuestionMarks	The number of question marks
Exclamation	The number of exclamations

B. Sentiment Classification Models

Eight different classification models (classifiers) were used: AdaBoost, LogitBoost, Bagging, SVM, logistic regression, Neural Networks, BayesNet, and decision tree [23][24]. Given the small number of observations in the training data and the limited number of features considered, it was possible to examine all model combinations of features for each classifier and find the feature set that best classified sentiment in the training set for each classifier. Classification accuracy and ROC area were used as measures of goodness of fit. Classification accuracy is the percentage of training data observations correctly classified. The ROC curve for a binary classifier system is a plot of the true positive rate vs. the false positive rate for varying discrimination thresholds. The ROC curve is a measure of the ability of a classifier to produce good relative instance scores, and is insensitive to changes in class distribution [25]. ROC area is simply the area under the ROC curve. High ROC area and high classification accuracy are characteristic of a good classifier. To avoid being too tied to the one training set, goodness of fit statistics are also estimated using 10-fold cross-validation.

The goodness of fit statistics for the best fitting feature set for each classifier are given in Table IV. AdaBoost, where regression trees are used as weak learners, had the best ROC Area (0.832) and classification accuracy (79.2%). This model used *PostLength*, *Neg*, *PosVsNeg*, *Name*, *Slang*, *PosStrength*, and *NegStrength* and had a false positive rate of 0.152, and false negative rate is 0.33. Not all available features were used in the final model. The AdaBoost classifier using all features has poorer fit, with ROC area of 0.813 and classification accuracy of 75.2% (see Table III) demonstrating what is often the case that too many features can reduce the prediction ability of the model. Note that the excluded features are not necessarily weak. For example, *Pos* by itself has high classification power (see column 2 of Table V) but is not included in the final model because it does not contribute additional discrimination given the other features already included (see column 4 of Table V).

TABLE IV
BEST FIT CLASSIFIERS, THEIR ROC AREA, AND CLASSIFICATION ACCURACY.

Classifier	ROC Area	% Classification Accuracy
AdaBoost	0.832	79.2%
Logistic Regression	0.832	77.5%
LogitBoost	0.816	76.8%
BayesNet	0.802	74.2%
Bagging	0.794	73.5%
Neural Networks	0.785	73.8%
Decision Tree	0.782	77.2%
SVM	0.658	75.2%

Considering the complexity of sentiment analysis, the observed performance of the classifiers is typical. Classification accuracy for other sentiment analyses reported in the literature for various domains range from 66% for movie reviews to 84% for automobile reviews [19][17].

C. Feature Analysis

Using AdaBoost with the best fitting feature set, further studies were preformed to establish the importance of each feature to model performance, and to explore how individual features differ between posts classified to negative or positive classes. The goal of this latter analysis was to answer questions such as: *Are negative posts less likely to mention users' names?* in which case the *Name* value should be less in negative posts, and *Can positive posts contain negative words?* assessed by looking at *Neg*.

A boxplot of features for negative and positive posts is given in Figure 4. This figure indicates that negative posts are slightly longer, contain more negative words, and have more negative strength on average, while positive posts mention more names, have higher ratios of positive and negative words, have larger positive strength on average, and use slightly more slang. Student's t-tests assuming separate variances were used to assess the importance of individual features. The differences for *PostLength* ($P = 0.08$), *Neg* ($P = 0.68$) and *Slang* ($P = 0.06$) are not significant, while *PosVsNeg* ($P < 0.001$) and *Name* ($P < 0.001$) are. T-tests were not run for *PosStrength* and *NegStrength* because they contain too few value points ($\{1, 2, 3, 4, 5\}$ and $\{-5, -4, -3, -2, -1\}$, respectively) and as a result do not satisfy Normality assumptions.

The importance of individual features in the best fitting feature set for AdaBoost is measured in two ways: first by examining the fit of the AdaBoost model with each feature used alone, and next, by examining whether adding the feature last into the model significantly improves overall model fit. Model fits for these two approaches are shown in Table V. Note that *PosStrength* and *PosVsNeg* are the individual features best able to classify correctly alone in the model but these features are not greatly better than a number of the other features. The last-in statistics shows the performance decrease from the best fitting AdaBoost classifier if individual features are dropped from the model. Leaving any feature in the best fitting feature set does show some performance deterioration,

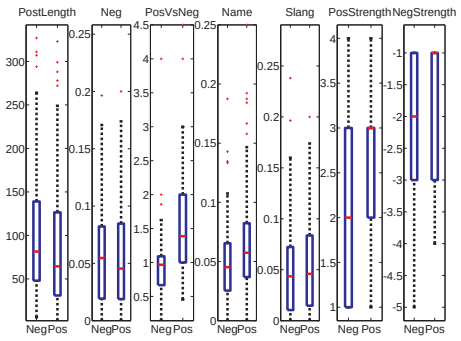

Fig. 4. Boxplots of features for negative (*Neg*) and positive (*Pos*) classes.

TABLE V
IMPORTANCE OF FEATURES ASSESSED INDIVIDUALLY OR AS LAST ADDED FOR THE ADABOST CLASSIFIER.

Feature introduced as:	Only ROC Area	Last-in ROC Area	Decrease from Best Fit
PosStrength	0.696	0.774	0.054
PosVsNeg	0.694	0.804	0.024
Neg	0.459	0.813	0.015
Slang	0.527	0.813	0.015
NegStrength	0.544	0.813	0.015
PostLength	0.545	0.819	0.009
Name	0.572	0.820	0.008

although only *PosStrength* and *PosVsNeg* show large impact.

IV. SENTIMENT DYNAMICS

The fitted AdaBoost model was used to establish the sentiment level for all (manually labeled and unlabeled) posts in the CSN forum. For any post P, the model generates the predicted probability $Pr(P)$ that the post belongs to the *positive* class. The corresponding probability for the *negative* class is computed as $1 - Pr(P)$. If $Pr(P) > 0.5$, post P is classified as *positive*, otherwise, it is labeled as *negative*. There is, of course, some error in this classification since the fitted model is not perfect (ROC area=0.83, correct classification rate= 80%), but fit is sufficient for the purposes of analyzing sentiment dynamics in the CSN forum. Even if post P is incorrectly classified, its $Pr(P)$ is likely to be close to 0.5 suggesting little confidence in the classification. $Pr(P)$ is used as a *Sentiment Indicator* that measures the likelihood that post P has *positive* sentiment in subsequent analyses.

For the 468,000 posts extracted from the CSN forum, using the fitted AdaBoost model and including the manually labeled posts, 45.9% of initial posts are classified as *negative* (54.1% classified *positive*) while 31.2% of all posts are classified as *negative* (68.8% classified *positive*).

A. Definition of Sentiment Change of Thread Originator

Sentiment change expressed by thread originators is measured as the difference between the *initial-sentiment* (denoted as $S_0(T)$) and the *subsequent-sentiment* (denoted $S_1(T)$) expressed within that thread (see Table VI). Threads that have no originator replies are excluded because it is not possible to estimate the change without a *subsequent-sentiment*. Threads without replies from anyone other than the thread originators are also excluded, because any sentiment change could not be attributed to interaction with others (impacting factors). With these exclusions, 23,164 or 47.5% of the initial 48,000 threads were left.

B. Sentiment Change of Thread Originator vs. The Number of Replies

The number of replies from others in a thread reflects the level of interest in the discussion topic. By definition, *negative thread originators* are those whose *initial-sentiment* are negative ($S_0 = -$) and *positive thread originators* are those whose *initial-sentiment* are positive ($S_0 = +$). A comparison of the number of replies to a thread by others (n_1) between *negative thread originators* and *positive thread originators* is shown in Figure 5. The curve marked by triangles shows the change in the likelihood that a *negative thread originator* ($S_0 = -$) changes to a positive sentiment ($S_1 = +$) as the number of replies from others increases. The curve marked by circles shows the likelihood that a *positive thread originator* ($S_0 = +$) stays positive ($S_1 = +$) as a function of the number of replies from others. About 75% of *negative thread originators* subsequently express positive sentiment when at least one reply from others is received, and the probability increases as the number of replies from others increases. A similar trend is shown for *positive thread originators*. For a given number of replies from others, the probability of positive sentiment from a *positive thread originator* is always greater than that of a *negative thread originator*. Intuitively, it is expected that *positive thread originators* are more likely to have positive *subsequent-sentiment* than *negative thread originators*. These curves do not reach 1.0 (100%), because regardless of the number of thread replies, some *negative thread originators* will never change to positive *subsequent-sentiment*, and not all *positive thread originators* keep a positive *subsequent sentiment*. This fact may be attributed to forum participants who are not satisfied with the interactions received. For example, conflicts occur in some threads, which lead to negative impacts on both participants' and originator *subsequent-sentiment*.

The number of self-replies (n_0) by a thread originator reflects personal involvement in the thread. Figure 6 shows the percent of *negative thread originators* (triangles) and *positive thread originators* (circles), who subsequently express positive *subsequent-sentiment* ($S_1 = +$) as a function of the numbers of self-replies to their own thread (n_0). Both functions increase with increasing number of self-replies suggesting that the continued involvement of originators in the thread they started is positively associated with *positive subsequent-sentiment*. Only a small number of threads have high numbers of self-replies leading to uncertainty in the true form of these functions. The dip in the negative thread originators function in the middle of the range and the drop in the positive thread originators function at the upper end of the range may be real, but may also reflect small-sample biases.

TABLE VI
DEFINITION OF *initial-sentiment* ($S_0(T)$) AND *subsequent-sentiment* ($S_1(T)$) OF A THREAD (T).

Variable	Value	
Initial-sentiment (S_0)	Positive (+)	If $Pr(P_0) > 0.5$, where P_0 is the initial post by the originator, and $Pr(P_0)$ is the probability output by the sentiment model
	Negative (-)	Otherwise
Subsequent-sentiment (S_1)	Positive (+)	If $\sum_{i=1}^{n_0} Pr(P_{0i})/n_0 > 0.5$, where P_{0i} are replies by the thread originator (self-replies), and n_0 is the number of self-replies
	Negative (-)	Otherwise

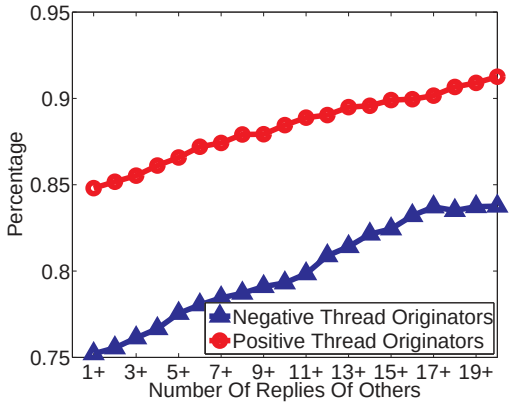

Fig. 5. Circle: $y = p(S_1 = +|S_0 = +, n_1 \geq x)$; Triangle: $y = p(S_1 = +|S_0 = -, n_1 \geq x)$, where n_1 is the number of replies from people other than the thread originator, S_0 and S_1 are *initial-sentiment* and *subsequent-sentiment*, respectively

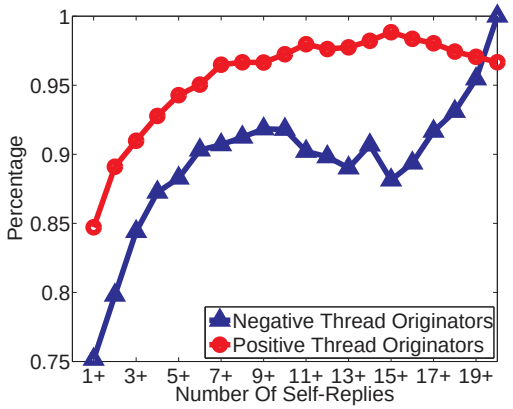

Fig. 6. Circle: $y = p(S_1 = +|S_0 = +, n_0 \geq x)$; Triangle: $y = p(S_1 = +|S_0 = -, n_1 \geq x)$, where n_0 is the number of self-replies by the thread originator, S_0 and S_1 are *initial-sentiment* and *subsequent-sentiment*.

C. Sentiment Change Indicator

Further insights regarding factors that may contribute to the change of sentiment of thread originators is obtained by analysis of the *Sentiment Change Indicator* for a thread originator defined as

$$\Delta_{Pr} = \sum_{i=1}^{n_0} Pr(P_{0i})/n_0 - Pr(P_0),$$

where P_0 is a thread originator's initial post in a thread, P_{0i}, $1 \leq i \leq n_0$, are self-replies by the thread originator, n_0 is the number of self-replies, and $Pr(.)$, generated by the AdaBoost-based sentiment model, indicates the probability of a post being classified as *positive*.

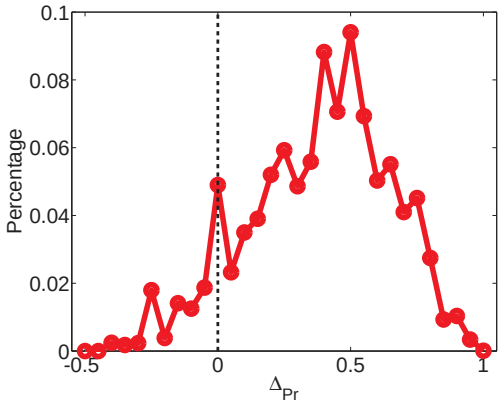

Fig. 7. Distribution of *Sentiment Change Indicator* (Δ_{Pr}).

The larger the Δ_{Pr}, the more likely that self-replies will be classified as positive by the sentiment model. Figure 7 is a rough estimate of the distribution of Δ_{Pr} for *negative thread originators*. From this distribution, roughly 7.9% of *negative thread originators* have $\Delta_{Pr} < 0$. The average Δ_{Pr} is 0.14 (standard deviation=0.31) which is significantly larger than 0, suggesting that after interacting with others, *negative thread originators* are likely to publish posts that are of *positive-sentiment*.

D. Factors related to the Sentiment Change Indicator

The time a participant spends replying to a thread is highly correlated to the average number of words used in their replies (a quantity referred to as the *average length of replies*). The association between *average length of replies* and Δ_{Pr} is explored in Figure 8. Each point is a thread, the black line is the fitted linear regression, and the circle points are average Δ_{Pr} values computed for all threads within a small "window" of *average length of replies* for "windows" defined from left to right on the *average length of replies* axis. The regression slope is not statistically different from zero suggesting no relationship of Δ_{Pr} with *average length of replies*. For replies greater than 11 words on average, $\Delta_{Pr} \approxeq 0.4$. For *average length of replies* below 11 words, the Δ_{Pr} declines. This suggests that for the most part, the average length of reply is not an influential factor in changing sentiment.

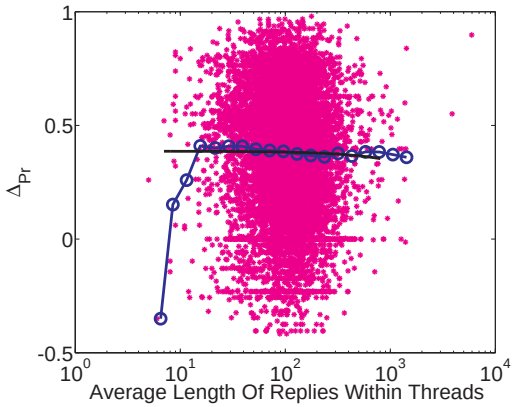

Fig. 8. *Average length of replies* is not a major driver of *Sentiment Change Indicator* Δ_{Pr}.

Intuitively, positive or negative sentiments may propagate, suggesting that the average sentiment of replies (Pr) by other people may contribute to the *Sentiment Change Indicator* (Δ_{Pr}). The *Sentiment Change Indicator* and average sentiment of replies for each thread are plotted as points in Figure 9. The slope of the linear regression (the straight line in the plot) of Pr on Δ_{Pr} is 0.125 and significantly different from zero ($P < 0.05$). The "window" average curve (circle points) confirms the linearity of the relationship. This relationship suggests that to some extent, higher average sentiment in subsequent non-originator posts increases the likelihood of a positive sentiment change in the thread originator. The broad spread of the points suggests that this may not be a major driving factor.

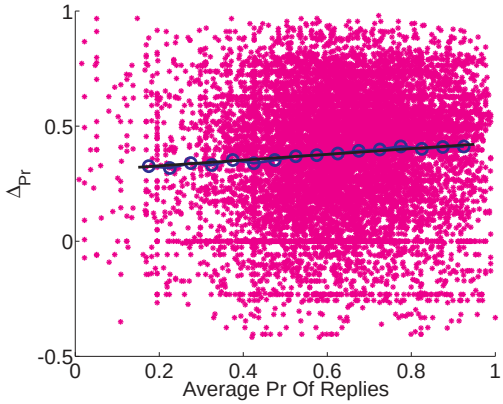

Fig. 9. Increasing *Sentiment Indicator* of replies by other people increases the originator's *Sentiment Change Indicator*.

The importance of the topic of a thread to the community can be measured by the time elapsed before the first reply by other people. A scatterplot of Δ_{Pr} and the time elapsed before the first reply by other people is shown in Figure 10. The slope of the linear regression is not statistically different from 0, and the linearity of the relationship is confirmed by the "window" line (circle points). There is indication that when the time interval before first reply is very small (less than about 3.5 hours or 210 minutes), the average Δ_{Pr} is slightly larger suggesting that only very timely replies may contribute to an increase of Δ_{Pr}.

Table VII summarizes the effects of the number of self-replies, the number of others' replies, the average number of words of others' replies, average sentiment of others' replies, and the time elapsed before the first reply by others.

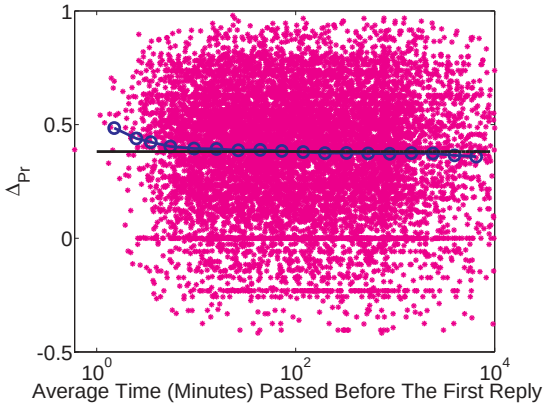

Fig. 10. *Sentiment Change Indicator* (Δ_{Pr}) vs. the time elapsed before the first reply

V. Conclusions and Future Work

Using posts in an online health community for cancer survivors and caregivers (the ACS's Cancer Survivors Network) containing a half million forum posts made over a period of ten years, sentiment analysis was conducted and the dynamics of online users' sentiment revealed. Using machine learning techniques to extract important features, multiple candidate classification models were studied from which an automated sentiment classifier was developed. The final AdaBoost sentiment model was used to identify post features useful in predicting participant sentiment. Finally, the sentiment dynamics of thread originators in discussion threads were analyzed and factors driving the dynamics studied.

This is the first study that researches the sentiment benefits and associated dynamics of a large-scale health-related electronic forum. Modeling and analysis showed that 75% to 85% of users in the CSN forum experienced a positive change in sentiment through their online interaction with other forum participants. The greater the number of replies by others to a thread, the more likely is the thread originator's subsequent sentiment to be positive, regardless of the sentiment in the initial post. Furthermore, the level of involvement of a thread originator in subsequent thread posts is positively correlated with their positive subsequent sentiment. Finally, the higher the average sentiment of other posts in a thread, the more likely the sentiment change of the thread originator will be positive.

In this work a classification is applied to each post as one entity, regardless of length. This post-level analysis has its

TABLE VII
FACTORS FOR SENTIMENT CHANGE OF THREAD ORIGINATORS

Factor	Effects
Involvement (self-replies)	The more involved, the more likely a thread originator expresses positive *subsequent-sentiment*
The number of replies by others	The more replies from others, the more likely a thread originator expresses positive *subsequent-sentiment*
The average number of words in others' replies	If all replies by others are too brief, then a thread originator is likely to express no or little positive change of sentiment
The average sentiment indicator (Pr) of others' replies	If replies by others are more positive, then a thread originator is likely to express more positive sentiment change
The time elapsed before the first reply by others	If the first reply from other people is very timely, then a thread originator is likely to express more positive sentiment change

limitation in that multiple sentiments about different topics are possible in one post, therefore, developing more fine-grained sentiment analysis, analogous to analyzing the sentiment about features of products [19][17], is an important direction for future research. While a binary classification of sentiment was used in this work (*positive* and *negative*), future studies will likely wish to fine tune sentiment using a multi-level classification system. This will require a different and more complex definition of a *sentiment change indicator*.

This study introduced two new features of posts, *Name* and *Slang*, which proved to be useful in discriminating *positive* from *negative* sentiment in posts. These features, along with others yet to be identified, need to be studied in other contexts to understand their real worth for assessing sentiment in online health communities. Other approaches to improving the sentiment classification model, such as subjectivity summarization [26] also need to be explored.

This research is significant in that it is a prototype others may find useful in guiding their analyses of the dynamics, impact, and opinions that may affect the emotion and the wellbeing of individuals and subgroups in other online health communities. In addition, this analysis has identified factors associated with *positive sentiment change* that designers of new online health communities and managers of existing online health communities may find useful. The ultimate goal of this research is producing online health communities that support the emotional health of community members and contributes to improved quality of life for those dealing with disease or supporting someone who is dealing with disease.

REFERENCES

[1] ACS, *American Cancer Society Cancer Facts and Figures 2011*, 2011.

[2] WHO, "Cancer," World Health Organization, Tech. Rep., 2009.

[3] J. Barraclough, *Cancer and Emotion: A Practical Guide to Psycho-oncology*, 3rd ed. Wiley, 1999.

[4] L. Eaton, "Europeans and americans turn to internet for health information," *British Medical Journal*, vol. 325, no. 7371, pp. 989–989, 2002.

[5] K. Zickuhr, "Generations 2010," Pew Internet, Tech. Rep., 2010.

[6] S. Ziebland, A. Chapple, C. Dumelow, J. Evans, S. Prinjha, and L. Rozmovits, "How the internet affects patients' experience of cancer: a qualitative study," *British Medical Journal*, vol. 328, no. 7439, p. 564, 2004.

[7] J. Preece, "Empathic communities: Balancing emotional and factual communication," *Interacting with computers*, vol. 12, no. 1, pp. 63–77, 1999.

[8] A. Bambina, *Online social support: the interplay of social networks and computer-mediated communication*. Youngstown, N.Y.: Cambria Press, 2007.

[9] C. Dunkel-Schetter, "Social support and cancer: Findings based on patient interviews and their implications," *Journal of Social Issues*, vol. 40, no. 4, pp. 77–98, 1984.

[10] S. Rodgers and Q. Chen, "Internet community group participation: Psychosocial benefits for women with breast cancer," *Journal of Computer-Mediated Communication*, vol. 10, no. 4, 2005.

[11] C. E. Beaudoin and C.-C. Tao, "Modeling the impact of online cancer resources on supporters of cancer patients," *New Media and Society*, vol. 10, no. 2, pp. 321–344, 2008.

[12] A. J. Winzelberg, C. Classen, G. W. Alpers, H. Roberts, C. Koopman, R. E. Adams, H. Ernst, P. Dev, and C. B. Taylor, "Evaluation of an internet support group for women with primary breast cancer," *Cancer*, vol. 97, no. 5, pp. 1164–1173, 2003.

[13] D. Maloney-Krichmar and J. Preece, "A multilevel analysis of sociability, usability, and community dynamics in an online health community," *ACM Trans. Comput.-Hum. Interact.*, vol. 12, no. 2, pp. 201–232, 2005.

[14] D. A. Redelmeier and D. Kahneman, "Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures," *Pain*, vol. 66, no. 1, pp. 3–8, 1996.

[15] M. S. Litwin and K. A. McGuigan, "Accuracy of recall in health-related quality-of-life assessment among men treated for prostate cancer," *Journal of Clinical Oncology*, vol. 17, no. 9, pp. 2882 – 2888, 1999.

[16] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, and M. Gutmann, "Life in the network: the coming age of computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[17] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2(2008), pp. 1–135, 2008.

[18] I. Yalom, *The Theory and Practice of Group Psychotherapy*. New York: Basic Books, 1970.

[19] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer, 2006.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[21] M. Hu and B. Liu, "Mining and summarizing customer reviews," *The Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2004)*, 2004.

[22] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[23] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*. Morgan Kaufmann, 2011.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[25] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *Machine Learning*, 2004.

[26] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," *The Proceedings of the Association for Computational Linguistics (ACL'2004)*, 2004.