Analyzing Patient Interactions within Cancer Support Groups

Zhenghao Chen, Pang Wei Koh
CS 224N Class Project
zhenghao,pangwei@cs.stanford.edu

1. Introduction

Cancer is the leading cause of death in the developed world, claiming the lives of approximately 7.6 million people in 2008 alone [1]. Cancer treatment, be it by surgery, chemotherapy, radiotherapy or any combination thereof, is often debilitating, and its effects can be felt for years even among patients who have been "successfully" treated for cancer. These treatments often lead to a host of chronic side effects, such as fatigue, insomnia and lymphedema. Moreover, the ever-present fear of recurrence among survivors frequently contributes to high levels of depression and stress [2].

Management of cancer survivors, especially in the early post-treatment stages, is therefore of paramount clinical importance. One common approach to cancer survivor management has been the establishment of peer support groups, which allow people battling the effects of cancer and its treatment to band together for mutual support and encouragement. More recently, purely internet-based peer support groups have emerged, offering a wealth of data in the form of recorded conversations and interactions between participants on discussion forums.

In this paper, we present a systematic way to extract information from the unstructured text data available in these online peer support groups. To the best of our knowledge, this has not been done before. Successful mining of this data would shed light on important questions such as:

(1) *Are online peer support groups useful for cancer survivors?* The verdict is still out on whether such peer support groups with minimal supervision from healthcare professionals are beneficial for their participants. While studies attempting to address this question have been conducted [3], these studies largely rely on quantitative data obtained through methods such as surveys and laboratory tests, which are difficult and expensive to perform. On the other hand, text data from discussion forums is a plentiful and relatively inexpensively obtained source of data which could also be used to answer this question. One simple metric of usefulness is to track the semantic orientation of posts - how positive or negative the posts are - over time: if they get steadily more positive, for example, one can infer that the course might be doing something beneficial. By studying the dynamics of text sentiment over the course, we aim to provide a complementary method of assessing the efficacy of online peer support groups.

(2) *Can we automatically discover better clinical practices from text data?* By analyzing the issues that participants are talking about on these peer support groups, we hope to identify new targets for clinical intervention. Such medical crowd-sourcing methods are extremely new, but recent studies [4, 5] indicate that they can yield fruitful results.

(3) *Can we predict health outcomes more accurately based on text data?* The ability to detect in real-time participants who are more likely to suffer more strongly from conditions such as depression, and highlight these participants to medical professionals would be of significant clinical importance. Analyzing text data (as opposed to self-reported survey scores) could potentially allow us to performing sentiment analysis at a higher level of granularity and investigate the relationship between emotional/social health and its link to downstream health status.

We will first describe our data and the general analysis methodology, before treating each of these questions in turn.

2. Methodology

We ran our experiments on Stanford University's "Cancer: Thriving and Surviving" self-management workshop, found at https://cancersurvivors.stanford.edu. Each workshop lasts 6 weeks, comprises approximately 25 participants and is facilitated by 2 trained moderators. The workshops involve online lessons and a discussion forum, where participants are free to post and reply to one another. The discussion forum was broadly structured into four categories: Problem-solving, Emotions, Celebrations and Actions. The latter category collected the weekly action plans that the participants were asked to complete (e.g. "Walk three times for 30 min this week."). Participants were all aged 19 and over, had been diagnosed with cancer within the last 5 years, and had also completed cancer treatment (with the exception of long-term hormonal therapy, which is especially common for breast and prostate cancer).

In this pilot study, we used data from 48 participants over 2 workshops, collaborating with the Patient Education Research Center in the Stanford School of Medicine. This data included more than 3000 forum posts and two questionnaires, one taken at the start of the course and the second 6 months after, which asked the participants a series of approximately 100 questions about their physical and mental well-being. Following the methods used in [6], we used these answers to obtain summary scores for general health, quality of life, fatigue levels, illness intrusiveness, depression, stress, degree of physical exercise, and so on. We also had access to a demographic survey, which asked for basic participant information like age, sex, race, type of cancer, time since end of treatment, and so on.

2.1. Sentiment Analysis

We want to identify the *sentiment polarity* of forum posts in order to identify topics related to positive or negative sentiment, as well as track the time evolution of participant sentiment through the 6 week course. Specifically, we are interested in performing a binary sentiment polarity classification task, where each post is labelled either positive or negative. The test data consisted of 2899 unlabelled forum posts from 48 participants. In addition, we separated and hand-labelled 303 other posts to serve as a validation/evaluation set.

2.1.1. Supervised Learning

We first investigated a supervised learning approach, following the methodology described in [13]. This was done by modifying our submission for PA3 to implement a maximum entropy classifier, which was then trained on a training set of 10662 snippets of movie reviews (5331 positive and 5331 negative) made available by Pang and Lee [15].

Standard Unigram Model. We obtained baseline performance by evaluating a standard unigram-presence feature set. To produce this feature set, we scanned the training corpus and removed all words which occured 6 times or less, as well as a pre-defined list of 531 stopwords [22]. This resulted in a vocabulary of 1382 words. Each data example was represented as a binary vector reflecting the presence or absence of these 1382 words. Additional features such as presence of special punctuation ("!" and "?") were included as additional binary features.

The standard unigram model achieved an accuracy of 68.3% on our hand-labelled validation set. Error analysis of the validation set as well as the unlabelled main test set showed that a significant portion of posts were misclassified due to spurious word associations. For example, a short post containing the sentence fragment "... scars that are hidden by my clothes..." was classified as positive because it contained the word "clothes", which had been associated with a positive sentiment. This problem was exacerbated by the problem of feature sparsity. As the language used in our training set (movie reviews), from which we derived our features, was qualitatively different from the language used in our test data (discussions about dealing with cancer), this led to context-specific and semantically important words in our test set (such as "scars") being ignored because they did not occur in our training set. In their place, emotionally unimportant words like "clothes", as shown above, were instead used as a spurious basis for classification.

High Valence Unigram Model. To minimize this overfitting, we observed that it was possible to determine the polarity of most posts by identifying just a few salient word-level indicators and their polarities. We hypothesized that restricting unigram features to a set of generic words with high valencies would make good features for our classifier, as it would control the problems of spurious associations described above. This motivated the high valence unigram model, in which each post is represented by a binary vector reflecting the presence or absence of 1436 hand-labelled common high valence words taken from [19].

By restricting our feature set to high valence words, we achieved a greatly improved accuracy of 73.9% on the validation set. However, while the use of high valence words as features allowed us to effectively classify a large number of posts, the relatively low number of such terms meant that the model had trouble disambiguating posts that did not contain any generic high valence terms. We also note that we can observe the effect of overfitting directly through the log-likelihood objective at convergence which was 7148 for the high valence model compared to 6615 for the standard unigram model. Therefore, while the standard unigram model is able to achieve a better objective value, indicating that it is modelling the training data better, the high valence model is able to generalize to our patient data better.

Standard Bigram Model. It has been commonly noted that unigram feature approaches are unable to account for context-specific polarity [15]. For example, while "problem" is a negative indicator word, the phrase "no problem" has positive polarity because of the "no" negation. This led us to investigate the use of bigram features for sentiment classification. We first implemented the standard bigram feature set, in which we considered only bigrams which occur 5 times or more in the training corpus. This standard bigram model achieved a validation accuracy of 64.4%. In our error analysis, we noted that performance was poor compared to the standard unigram model, since the gain in context came at the cost of greatly increasing the number of features (from 1382 to 4678), which in turn exacerbated the problems of overfitting and data sparsity.

High Valence Bigram Model. As with the unigram models, we noted that the concept of only using high valence indicator words could be extended to bigram features by only considering bigrams in which at least 1 word was a high valence word. While selecting for high valence bigrams reduced the number of features and helps prevent overfitting, features remained sparse, with only 588 high valence bigrams occuring in the training corpus more than 4 times. The high valence bigram model achieved a validation accuracy of 41.6%, a poor performance which we determined was due to the extreme sparsity of high valence bigrams in both training and validation data.

Error Analysis. As alluded to above, the main problem we faced with a supervised learning approach was a lack of generalizable training data. Due to the fairly specialized nature of our problem domain, the overlap of sentimentally significant terms between our training and test set was minimal. Unfortunately, existing sources of supervised data with labelled sentiments consist mostly of reviews of products or services (movies, restaurants, electronics), all of which utilize a qualitatively different language style and vocabulary.

Furthermore, while a certain proportion of posts can be classified according to the presence or absence of common high valence words/terms, there exist a fairly large subset of domain specific terms which have strong sentiment implications which are not captured in any generic list. Examples of such terms in our problem domain include negative indicators like "lymphedema", "mastectomy" and "Tamoxifen", and positive indicators like "remission" and "clear". Such terms would help to disambiguate posts which do not contain any clear generic indicator words.

Moreover, terms may have different connotations given the specific problem context. For example, the word "money" is typically a positive term but becomes a negative term in the context of our messages, since a common theme deals with the economic impact of cancer treatment. Similarly, in general usage, "cancer" is a typically negative term but has a more neutral connotation in the context of a cancer support group, since the majority of posts are concerned with the disease and can be positive (e.g., "I am coping well with cancer.").

We conclude that it is difficult to train on and generalize from available labelled sentiment analysis datasets to our specific problem domain of online, disease-specific support group data. This problem of poor generalization manifests itself as feature sparsity and overfitting. While it is possible to achieve decent performance using only generic words with high valence, it is unclear how to further improve performance by utilizing domain specific terms as indicators.

2.1.2. **Semi-supervised Learning**

The limited success of the supervised approach and the lack of suitable labelled training data motivated us to consider semi-supervised methods of sentiment analysis. In particular, we desire a method which is able to automatically discover and incorporate domain specific terms as well as determine the influence such terms should have on the semantic orientation of a text, while only using a minimum of explicit labelled data.

We utilized a variant of Turney's algorithm for semi-supervised semantic orientation classfication to achieve both aims [18, 16]. The only supervision in this method comes from defining a set of positive and negative reference words which are broadly generalizable indicators of positive or negative sentiment. In our implementation, we used the same AFINN-96 list of high valence words as above [19] . The semantic orientation of a novel word is then estimated as the comparative similarity of the word to our group of positive/negative reference words.

Following the methodology described in [18], the similarity between a word and a reference word is determined using the Pointwise Mutual Information (PMI), where the PMI between words 1 and 2 is defined as

$$PMI(word1, word2) = log\left(\frac{p(word1, word2)}{p(word1)p(word2)}\right),$$

and where $p(word1, word2)$ is the probability that word 1 and 2 co-occur in the same post. In our implementation, the relevant probabilities are replaced by the empirical estimates of these probabilites from our test set. The PMI is used to calculate a variant of Turney's semantic orientation (SO) score [16, 17] for each word:

$$SO(word) = \sum_{pos \in positive} SO(pos) * PMI(word, pos) - \sum_{neg \in positive} SO(neg) * PMI(word, neg)$$

Here, positive and negative refer to the set of positive and negative reference words respectively. The SO score of each reference word is pre-initialized to the numerical scores provided in the AFINN-96 list [19], with each score correspondingly approximately to how strongly a reference word indicates a positive semantic orientation. We can thus induce an extended and domain specific lexicon of "polar" words beginning from a seed set of generic high valence words.

Estimating the sentiment polarity of a given post is then done by simply summing the semantic orientation scores of the words in the post and looking at the total score of each

post. A post is considered positive if its total score is positive and negative if its score is negative. We note that the maximum entropy classifier is a linear classifier which learns the weights to assign to the presence or absence of word features. These weights correspond to how strongly each word determines semantic orientation. Therefore, the summation method described here can be viewed as implicitly learning a classifier with our domain specific terms as the features.

High Valence Weighted Sum. In order to demonstrate the importance of having domain specific indicator terms, we attempted to classify posts *without* extending the provided seed reference list. This corresponds roughly to the supervised high variance unigram (frequency) model where the weights of each feature are coarsely pre-determined. By summing the scores of high valence words within each post and classifying them, we achieved an accuracy of 58.7% on the validation set. Analysis of the validation set showed that many posts had a score of 0, as these posts did not include any of the words in the (relatively small) set of reference words. However, some of these supposedly neutral posts contained strong domain specific indicators, for example, "...I've been on *Tamoxifen* I don't feel I think or remember as well as I used to." This shows that performance can be improved by learning a larger PMI induced lexicon.

PMI-Induced Lexicon. We then implemented the semantic orientation scoring function descibed above. SO scores were assigned to words which occured in the test set more than 4 times and co-occured with a reference word more than 4 times. SO scores for domain specific indicators were scaled to smaller magnitudes to reflect our intuitions that such scores were slightly less reliable than the original reference scores. As part of analysis and as a sanity check, we also looked at indicator terms we already knew to see if the scores assigned were consistent with our expectations.

As expected, terms such as "chemo" (-1.976), "surgery" (-1.203) and "terrified" (-0.9128) recieved low scores, indicating that they were strongly negative terms. Conversely, terms such as "cats" (2.189), "loved" (2.030) and "physician" (1.306) were predicted to be positive terms. Interestingly, we see signs of "overfitting" in the unsupervised context as well. For example, "stationary" (-0.9586) has a negative score which is not easily understood until one realizes from reading the actual posts that there are many complaints about "stationary bikes" in the data; likewise, "bike"s (-1.058) and "treadmills" (-1.013) are also negatively associated. Technical terms such as "lymphedema" (-1.07), "diagnosed" (-1.9822), prostate (-1.3533) and "MRI" (-0.513) were also found to be indicative of semantic orientation, which is notable since such specific terms would not have surfaced using training data from a different domain in a supervised setting.

Using the PMI-induced lexicon, validation accuracy increased to 77.6%, surpassing the best supervised learning approach (high valence unigram model).

Zoning and Negation Detection. As mentioned above, the inherent disadvantage of unigram approaches is the difficulty in incorporating context. For example, although co-occurence of terms is defined as occuring in the same post, we noticed that there was a tendency for certain long posts to contain both positive and negative statements. A negative statement would also often be followed by a comforting thought or a consolation which would contain positive reference terms. We therefore wanted to incorporate our intuition that changes in semantic orientation tend to occur, but relatively slowly. Therefore, if two terms co-occur and are located close together within a post, then there is a larger probability that they have similar semantic orientations; conversely, terms which co-occur far away from each other are less likely to share semantic orientation. We thus introduced a "zoning" weighting system in which co-occurences of terms in closer proximity have a relatively higher weightage in the calculation of the PMI. As expected, zoning made misclassified terms such as "mastectomy" (2.37 to 0.82) shift towards the appropriate orientation.

Also, as noted previously, changes in semantic orientation are often indicated by word cues such as such as "not" or "although". Indicator terms which have been modified by a negation should contribute the negative of their usual score to the overall post orientation. Negated words are thus important potential sources of misclassification, since the effect of

Learning Algorithm	**Validation Accuracy**
Supervised Learning	
Standard Unigram	68.3%
High Valence Unigram	73.9%
Standard Bigram	64.4%
High Valence Bigram	41.6%
Semi-supervised Learning	
High Valence Weighted Sum	58.7%
PMI-Induced Lexicon	77.6%
PMI-Induced Lexicon with Zoning and Negation Detection	**79.2**%

TABLE 1. Comparative Performance of Sentiment Analysis Algorithms

ignoring negation is effectively double the SO score of the negated word. We utilized a regular expression based negation detection algorithm since it was shown by Gxoryachev et el. [20] to outperform other methods such as classifier based approaches. Specifically, the NegEx negation detection library developed by Solti et al. [21] was used to detect negated indicators and flip the indicated orientation of such words. Our heuristic attempts to account for context further boosted our performance on the validation set up to 79.2%.

And the winner is... As summarized in Table 1, we find that unsupervised methods outperform supervised methods in our case, as domain appropriate training sets are not available. Furthermore, we show that a PMI-based induction approach is useful in determining domain specific indicators of semantic orientation. We use the PMI-induced lexicon with zoning and negation detection for the remainder of the experiments in this paper.

2.2. **Topic Modeling**

To achieve our goals of automatically discovering better clinical practices and predicting health outcomes from the data, we used a topic model to discover the underlying topic structure in our corpus. Specifically, we applied Latent Dirichlet Allocation (LDA), a generative model that models each document as a mixture of topics [7]. In the LDA model, each document draws a multinomial θ over topics from a Dirichlet distribution α. Each word in the document is generated by first drawing a topic z_i from the multinomial θ, and then drawing a word from the topic-specific multinomial β_{z_i} over words.

Our general idea was that we could inspect different parameters in the fitted LDA model to understand the underlying topic structure in our document. For instance, we could inspect the top words in each topic-specific multinomial β_z to understand the main topics of interest in our corpus, and aggregate statistics about which people talk about which topics and whether this correlates to any of the health outcome variables we are interested in. With this in mind, we tried applying LDA in different ways in order to get a model of the topic structure most consistent with these goals.

2.2.1. **Vanilla LDA**

First, we ran the plain LDA algorithm in MALLET on our corpus [8]. We treated each post as a separate document and tried with varying numbers of topics. Qualitatively, the topics that came out looked quite promising. For example, with 100 topics, the top 15 words were generally coherent and relevant to a single topic. We could easily identify topics that were related to food, sleep, goals, celebrations, exercise, family, cancer treatments, financial situations, etc. For example, Figure 2.1 gives a selection of food related topics we found in our topic model.

While these topics looked qualitatively promising, the end goal of the topic modelling was still to relate the topics to sentiments and downstream health outcomes. Unfortunately, LDA does not seem to automatically learn the distinction between people talking positively about a topic and people talking negatively about a topic. For example, consider the food related topics in figure 2.1. Each topic is clearly a topic about food, but it is unclear what

Top 15 Words
recipes veggies cooking cook broccoli tofu stir oil mixed juice toss tomato products fresh vegetable
eat eating food healthy vegetables diet veggies sugar fruit fruits snacks foods snack idea lunch
serving plate chicken orzo lemon size tender broth cups dill measure cup accomplishments diameter
morning tea cup coffee breakfast drink caffeine milk skip glass drinking haven open friendly read

FIGURE 2.1. Selection of food related topics from a 100 topic model of our corpus.

exactly the difference between the four different food topics is. Possibly the topics are split according to different types of food. It looks like the first topic is about vegetables, the second topic is about fruits and vegetables, the third topic is about recipes, and the fourth topic is about breakfast. However, we are more interested in healthy vs. unhealthy or positive vs. negative comments about food than we are interested in segregating different types of food. It may be the case that each of these topics has an even proportion of healthy food comments and unhealthy food comments. If this were the case, then counts of each of these topics would not help at all in predicting healthy diets. Other topics in our model show similar problems. As such, we explored using supervised topic modelling to obtain topics explicitly geared towards such tasks.

2.2.2. **Supervised LDA**

We could improve the saliency of the topics and possibly the predictive power of the topics if we forced the topics to correspond to the different variables we are looking for. To do this, we used Supervised Latent Dirichlet Allocation (sLDA), an extention to LDA that uses information from post labels in order to learn better topics [9]. sLDA essentially adds a new "response variable" y to each document, with y corresponding to the variable which we want to associate with the topics. The idea is that if we label the posts with their response variables, then the learning algorithm will learn topics that best predict the response variables.

We ran the following sLDA experiment for each variable of interest. First, we divided the patients into patients with this variable above the median and into patients with this variable above the median. Next, we labeled each post as “above median” or “below median” depending on if the author was above or below median on the variable of interest. Then, we trained an sLDA model with 100 topics using code from [10]. To visualize the results, we plotted the top 10 and bottom 10 topics according to their learned logistic regression parameter, similar to the visualizations used in [**?**].

One obvious problem from these plots was that there were many words common to all topics. We believe this is because the sLDA implementation we used uses a constant symmetric Dirichlet prior on the topic distributions. This is a known problem, as explored in [**?**]. In particular, the model does not want to create one topic that occurs much more frequently than other topics. But in our data there are topics that occur much more frequently than others, such as topics relating to cancer. To fix this problem, we modified the sLDA implementation to also run EM with respect to the Dirichlet parameter α, using the linear time Newton method described in [7].

Unfortunately, the results were still qualitatively poor (Figure 2.2). When we are running sLDA against the exercise variable, for example, all posts by high exercisers get labeled as “high exercise” and all posts by low exercisers get labeled as “low exercise.” However, most posts, even by high exercisers, are not related to exercise in any way at all, confusing sLDA when it tries to predict the exercise response variable based on posts that have nothing to do with exercise.

To see if we could fix this problem, we tried to treat each patient as one document, training an sLDA model on a corpus where we had concatenated each patients’ posts. However, this

approach did not work well either, as the concatenation of all posts together resulted in a big loss of information: the main way LDA learns about topics is by assuming that each document is made up of one consistent distribution of topics, which is no longer true when we combine all of one patients' posts.

One last attempt we made with sLDA was to see if sLDA could predict the post sentiment from our sentiment analysis after we removed high valence words from the posts. The idea is that sLDA might have been learning a topic full of positive words and a topic full of negative words simply because they correlated well with the response variables, while we wanted to learn which actual topics (such as exercise, insomnia, etc.) people were happy about. Figure 2.3 shows the results. As can be seen from this figure, sLDA still did not work very well. There is nothing clearly positive about the topics classified as positive, and nothing clearly negative about the topics classified as negative.

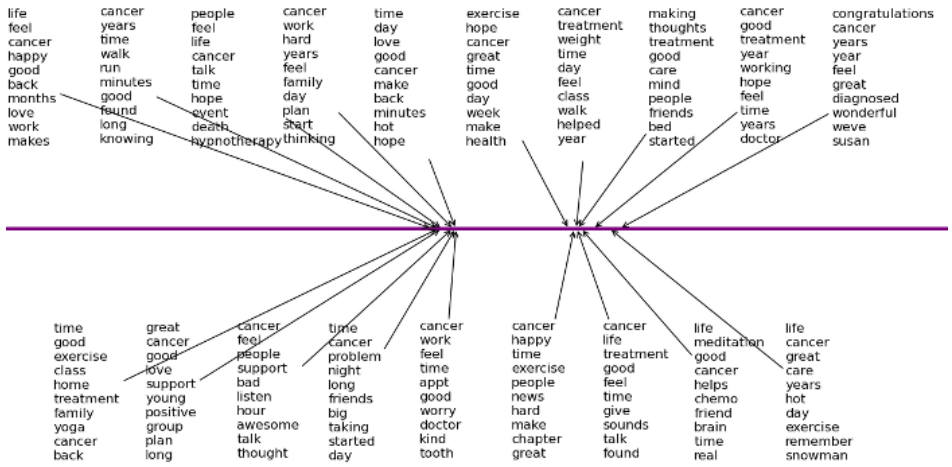


FIGURE 2.2. Top 10 topics correlated with high depression and top 10 topics correlated with low depression, from sLDA. Left is low depression, right is high depression.

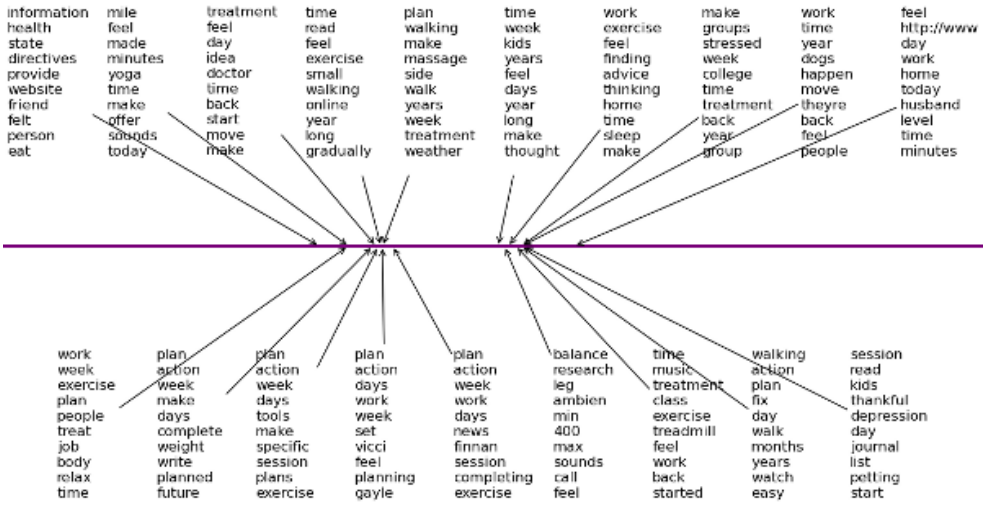
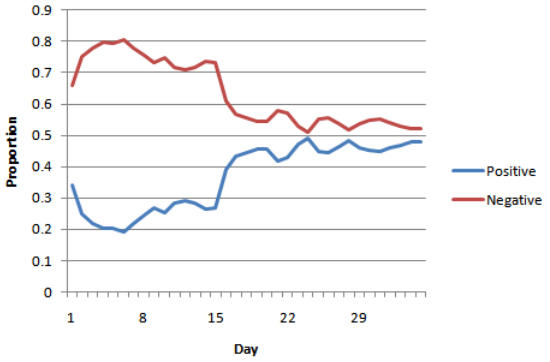


FIGURE 2.3. Top 10 topics correlated with positive sentiment and top 10 topics correlated with negative sentiment, from sLDA. Left is negative sentiment, right is positive sentiment.

FIGURE 3.1. Proportion of positive / negative posts over time, aggregated over patients and smoothed over 2 weeks



And the winner is... As seen above, supervised LDA did not work well on our limited data set, so we used unsupervised LDA for the remainder of the experiments in this paper.

3. Experiments and Results

3.1. Time Evolution of Sentiment

In order to analyze the direct impact that the online workshops had on their participants, we tracked the evolution of text sentiment over the 6 weeks of each workshop.

We observed that the first post of each forum topic was much more likely to be negative than the subsequent replies to that post, because the first post would often be a complaint or statement of a problem or some sort, while replies were almost invariably sympathetic and encouraging. This would throw off attempts to track sentiment evolution accurately; for example, particularly negative posts would often elicit a large volume of positive replies, so much so that the net sentiment contribution of the negative post would be overwhelmingly positive.

To obtain a more accurate picture of the emotional state of the participants, we therefore isolated all of the topic-starting posts and tracked only the evolution of text sentiment in these posts. This was done by classifying each post as positive or negative, as in Section 2.1, counting the proportion of positive vs. negative posts in each day, and then smoothing this over a 2-week sliding window (Figure 3.1).

We can see that as the course progressed, the proportion of positive posts steadily increased, while the proportion of negative posts dropped correspondingly. While there could be many possible confounding factors at play here (for example, participants might warm up to one another over the course of the workshop, resulting in steadily more positive posts), this result nevertheless suggests that the online support group was beneficial for the participants, at least in terms of how positive their online interactions became over time.

3.2. Associating Topics with Sentiments

In this experiment, we tried to discover which topics participants tended to write positively about and which topics they wrote negatively about. By running LDA on all of the posts in our corpus, we assigned each word in each post to one of 20 discovered topics. Subsequently, we ran sentiment analysis on the first posts in each forum topic, as in Section 3.1, and classified each post as positive or negative. For each topic, we took all of the words assigned to it and counted how many were in positive posts vs. negative posts. This gave us a measure of how often a particular topic was spoken of positively and negatively.

As expected, the topics that were most correlated with negativity concerned medical problems and side effects. The top words associated with the most negative topic were:

cancer treatment years year breast chemo back recurrence months treatments pain diagnosis diagnosed oncologist ll finished dx told doctor.

This matches well with our own observation that participants were the most stressed out and worried about possible recurrences and about their scheduled checkups at the oncologist / doctors, since each checkup would typically involve extensive testing and there was the possibility of a detected recurrence at each checkup. The next most negative topic was associated with:

side therapy scan don blood scars lymphedema radiation onc scar effects arm surgeon results physical reconstruction follow doctor pain.

This topic was concerned with the side effects of treatment, such as lymphedema (swelling due to poor lymph drainage) and the desire for physical reconstruction after mastectomy. As with the previous topic, there were a large number of posts involving such side effects in the corpus. While these two topics are not unexpected, they provide an important validation of our methodology.

The third most negative topic that we found was associated with:

sleep bed night work hours sleeping nights stressed trouble earlier early late schedule tired morning ready music problems times

This was less expected, and clinically relevant; it suggested that a large source of negativity was coming from problems with insomnia and sleeping habits. Once again, we corroborated this finding by manually examining the actual posts. The clinical-side implications of this finding is that a stronger focus on patient education about good sleep hygiene and ways to combat insomnia could possibly have large payoffs in terms of patient quality of life.

On the other side of the spectrum, the topic that was most correlated with positivity was associated with:

love kids year mom cat joy dogs years cats husband son home watching funny place sound house christmas watch,

reflecting the consistently happy posts about spending time with family, children and pets that were fairly abundant in our corpus.

More surprisingly, the third most positive topic was associated with:

plan action week peggy great days tools plans specific make session exercise walking confidence level completing time good complete.

This suggests that the participants generally felt positive about the weekly action plans that they were tasked to complete. Examination of the actual data validated this hypothesis: participants were uniformly happy when they managed to successfully complete an action plan, which happened more often than not, and at the end of the course they largely expressed satisfaction that they had learnt to make small goals and fulfil them. Moreover, action plans typically involved targets for exercise and walking, which are strongly associated with this topic. This result is an independent evaluation of the efficacy of having weekly action plans in these online workshops.

3.3. **Prediction of 6-Month Depression Levels**

We also investigated if the patient interaction data could be used as a predictor of health outcomes 6 months after the start of the online workshop. In particular, we focused on predicting self-reported depression levels, as they have been shown to correlate with prognosis in patients with breast cancer [2]. The aim of this experiment was to identify participants at higher risk of depression (defined by being above the median in terms of self-reported PHQ depression scores [12]). For this experiment, we worked with 40 participants, as 8 of them had not returned the 6-month questionnaire. We first labelled the participants based

on whether they were above or below the median in terms of their self-reported PHQ depression scores at 6 months, and then classified them using a linear SVM [11] with the $L2$ regularization parameter chosen by cross-validation.

To establish a baseline, we classified the participants based only on their actual PHQ depression scores at the start of the workshop and obtained a leave-one-out cross-validation accuracy of 62.5%. We then augmented the feature set with the total number of positive posts and the total number of negative posts that each participant wrote in the course. This resulted in a 5% increase in cross-validation accuracy to 67.5%.

We repeated this experiment on the *change* in PHQ depression from the start of the workshop to 6 months after. As with before, we attempted to predict which participants would be above the median in terms of the relative change of PHQ depression scores. Adding in the number of positive and negative posts resulted in an increase in cross-validation accuracy from 77.5% to 80%. These results show that even simple measures, like the total number of positive and negative posts, can result in slightly better predictive power.

Next, we explored if using the topic models learnt in the previous section could help in prediction. Specifically, we used a 20-topic model and counted the number of times each patient mentioned each topic, and used this to augment the feature set for each participant. It seemed like this count could be correlated to depression levels; for example, participants who talk a lot about the side effects from their therapy might stand higher risks of being depressed. Possible correlations also exist for other health outcomes: for example, participants who talk about food a lot might have a very healthy diet because they enjoy making healthy food, or perhaps a very unhealthy diet because they are consistently talking about their eating problems.

Unfortunately, these topic counts did not improve the accuracy of our prediction. Close examination of the errors made showed that the model was grossly overfitting, since it had 21 features (baseline PHQ depression score plus the counts of 20 topics) but only 39 training examples when performing leave-one-out cross-validation.

Moreover, not all of the topics that we found were as clean-cut as the examples shown in Section 3.2. For example, one topic was associated with:

hot flashes beautiful weather cold fall clouds trees tamoxifen fan cool wearing leaves sun winter summer base heavy sky

This topic contained both strongly positive words like “beautiful weather”, “clouds” and “trees”, and strongly negative words like “tamoxifen” and “hot flashes”, which would have confused the classifier. We discuss the implications of this below.

4. Discussion and Future Work

Our first two experiments (Sections 3.1, 3.2) show that patient interaction data can be used to evaluate both the online workshop in its entirety as well as individual components, such as the weekly action plans. Moreover, this data can be used to identify promising targets for therapeutic action, such as insomnia. While much of this work is highly speculative, it has three distinct advantages over the use of questionnaires and surveys. First, there might be an implicit selection bias in questionnaire data, as these questionnaires are often mail-in, and it is plausible that participants who did not care for the program much would be less likely to respond. Second, generic questionnaires might miss out pertinent questions if little is known about how the participants have actually been responding. Third, questionnaires are costly, especially in terms of the time and energy needed on both the participants’ and the clinicians’ sides.

The automatic data mining scheme that we present here is significantly more scalable to the larger target audience that these online support groups can attract. The use of this patient interaction data can thus serve to identify several preliminary but promising hypotheses (such as “weekly action plans benefit the participants the most”, or “stepping up efforts to combat insomnia will have a large positive impact”), before conducting more expensive and specific experiments to further investigate them.

Meanwhile, the experiment on predicting health outcomes (Section 3.3) demonstrates that it is possible to use the information in the interaction data to make better predictive models for the participants' health outcomes. This has two downstream applications: first, it allows for the identification of patients at high risk for social/psychological issues, which is hard to do from clinical tests. More targeted intervention, say in the form of professional counselling, could then be effected. Second, examining the features that contribute to the identification of risk could potentially shed light on pertinent clinical steps to take to minimize this risk.

We now discuss some of the more promising future directions of this work. First, as mentioned in Section 2.2, unsupervised LDA is completely agnostic to the variables we wish to associate the topics with, such as sentiment or PHQ depression scores. This gives rise to mixed topics, as in Section and 3.3, which would be detrimental for both topic-sentiment association and health outcome prediction. With more data, our hope is that supervised topic modelling (such as SLDA, described in Section 2.2) will allow us to design topics that are better suited for these tasks.

To aid in our own analysis, we read through each of the 3000+ forum posts and manually annotated the participants with a summary of relevant information, such as "Going through divorce. Proactive about exercise and weight management." This led to the observation that several drastic changes in health outcomes, in terms of depression, stress or illness intrusiveness, could be attributed to one-off big events. Even in our small dataset of 48 participants, we saw events like getting a job or being declared cancer-free corresponding to marked increases in happiness, while events like divorce and recurrence led to correspondingly significant drops. This suggests that a simple sentiment detection scheme that relies on word valencies will be inadequate for fully capturing the shifts in health outcomes over 6 months. More sophisticated sentiment analysis, possibly making use of domain-specific supervision to detect these big events, is one promising area of future research. Other ways to improve the sentiment detection scheme include augmenting the unsupervised approach to induce an extended lexicon over both unigrams and bigrams at the same time, as well as performing parts-of-speech tagging to isolate terms that carry more sentiment information, such as adjective-noun or adverb-verb phrases. Trying to classify and remove objective terms before performing sentiment analysis might also improve sentiment classification, as suggested in [14].

Obtaining and processing the required data was a major challenge in this project, in part due to the legal and ethical guidelines around patient confidentiality, and in part as analysis on this type of patient interaction data has not been performed before. As acquiring the data was a significant source of work on both sides, we limited ourselves to only studying participants from 2 workshops as a pilot project. This data sparsity and the resulting susceptibility to overfitting meant that we could not run more sophisticated experiments (attempting to do finer-grained identification of participants at risk of depression, for instance) with more features. With the simple proof-of-concept results documented in this paper, we are confident that we can move forward with analyzing significantly larger quantities of patient interaction data.

5. Conclusion

We have shown that patient interaction data in the form of unstructured text is a viable and promising source of information. While it has yet to be tapped, we believe that these results strongly hint that there is a useful signal contained within the text, and we see this paper as a preliminary step towards the full use of such unstructured text in informing medical decision-making.

Acknowledgements

This paper was written together with Marc Rasi. We are deeply grateful to Suchi Saria and Daphne Koller from the Stanford Computer Science Department, and Katy Plant, Philip Ritter and Kate Lorig from the Stanford School of Medicine for all of their assistance and advice.

References

[1] Jemal, A; Bray, F, Center, MM, Ferlay, J, Ward, E, Forman, D. "Global cancer statistics." CA: a cancer journal for clinicians. (2011)

[2] Hjerl K, Andersen EW, Keiding N, Mouridsen HT, Mortensen PB, Jørgensen T. "Depression as a prognostic factor for breast cancer mortality." Psychosomatics. (2003)

[3] Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, Anita Stern. "Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions." British Medical Journal. (2008)

[4] Wicks P, Frost J. "ALS patients request more information about cognitive symptoms." Eur J Neurol. (2008)

[5] Wicks P. "Excessive yawning is common in the bulbar-onset form of ALS." Acta Psychiatrica Scandinavica. (2007)

[6] Lorig KR, Sobel DS, Stewart AL, Brown BW, Bandura A, Ritter P, Gonzalez VM, Laurent DD, Holman HR. "Evidence Suggesting That a Chronic Disease Self-Management Program Can Improve Health Status While Reducing Hospitalization: A Randomized Trial." Medical Care. (1999)

[7] Blei, David M. and Ng, Andrew Y. and Jordan, Michael I. "Latent dirichlet allocation." J. Mach. Learn. Res. (2003)

[8] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. (2002)

[9] Blei, David M. and McAuliffe J. "Supervised topic models." NIPS. (2007)

[10] Finn Årup Nielsen. "AFINN word list." http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=5981

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. "LIBLINEAR: A Library for Large Linear Classification." Journal of Machine Learning Research 9(2008).

[12] Kroenke K, Spitzer R L, Williams J B. "The PHQ-9: validity of a brief depression severity measure." Journal of General Internal Medicine. (2001)

[13] Pang B, Lee L, Vaithyanathan S. "Thumbs up? Sentiment Classification using Machine Learning Techniques" Proceedings of EMNLP (2002)

[14] Pang B, Lee L. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts"Proceedings of the Annual Meeting of the ACL (2004)

[15] Pang B, Lee L. "Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales", Proceedings of the Annual Meeting of the ACL (2005)

[16] Turney P, Littman M. "Measuring Praise and Criticism: Inference of Semantic Orientation from Association" ACM Transactions on Information Systems, 21(4):315-346

[17] Rao D, Ravichandran D. "Semi Supervised Polarity Lexicon Induction" Proceedings of the European Chapter of the ACL (2009)

[18] Turney P. "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews" Proceedings of the Annual Meeting of the ACL (2002)

[19] Nielsen F. "AFINN-96" Department of Informatics and Mathematical Modelling, Technical University of Denmark (2010)

[20] Goryachev S, Sordo M, Zeng Q T, Ngo L. "Implementation and Evaluation of Four Different Methods of Negation Detection". Technical Report DSG (2006)

[21] Solti I, Kye J. "GenNegEx", http://code.google.com/p/negex/ (2010)

[22] Lewis D, Yang Y, Rose T G, Li F. "RCV1: A New Benchmark Collection for Text Categorization Research", Journal of Machine Learning Research (2004)