Homophily of Vocabulary Usage: Beneficial Effects of Vocabulary Similarity on Online Health Communities Participation

Albert Park, PhD[1], Andrea L. Hartzler, PhD[2], Jina Huh, PhD[3], David W. McDonald, PhD[4], Wanda Pratt, PhD[1,5]
[1]Biomedical Informatics & Medical Education; [4]Human Centered Design & Engineering; [5]Information School, University of Washington, Seattle, WA
[2]Group Health Research Institute, Seattle, WA
[3]Department of Media and Information, Michigan State University, East Lansing, MI

ABSTRACT

Online health communities provide popular platforms for individuals to exchange psychosocial support and form ties. Although regular active participation (i.e., posting to interact with other members) in online health communities can provide important benefits, sustained active participation remains challenging for these communities. *Leveraging previous literature on homophily (i.e., "love of those who are like themselves"),* we examined the relationship between vocabulary similarity (i.e., *homophily of word usage)* of thread posts and *members'* future interaction in online health communities. We quantitatively measured vocabulary similarity by calculating, in a vector space model, cosine similarity between the original post and the first reply in 20,499 threads. Our findings across five online health communities suggest that vocabulary similarity is a significant predictor of *members' future inter*action in online health communities. These findings carry practical implications for facilitating and sustaining online community participation through beneficial effects of homophily in the vocabulary of essential peer support.

INTRODUCTION

Many people use online health communities, such as WebMD and Facebook health groups, to exchange peer support and connect with others[1]. Research on the benefits of online health communities highlights psychosocial benefits—such as reduced depression[2,3] and stress[4,5]—from active participation in online health communities. However, sustaining active participation remains a prominent challenge for online communities in general[6–10] due to issues like lurking (i.e., participating without posting) and dropouts.

Sustained, active participation in online communities has been shown to positively correlate with a number of different factors. For example, receiving a response to a newcomers' first post[9], receiving emotional support[8], obtaining a sense of community[11], and having familiarity with online interactive services (e.g., chat)[6] have all shown to positively correlate with active participation or degree of effort and time spent with the community. Although these studies provide insight on how to sustain active participation in general online communities, only one study examined online health communities[8]. In contrast to the typical online community, active participation in online health communities could have implications for quality of life due to the purpose of participation: exchanging health information and psychosocial support. In that study of an online health community[8], participation was measured by sign-ins, which includes passive behaviors like lurking[10]. Lurkers not only gain fewer benefits than active participants[4] but they also do nothing to promote community sustainability. In our study, we focus on active participation to better reflect the benefits of online health communities to members and community sustainability.

Homophily, the tendency for individuals to be attracted to others with similar characteristics such as attitude and behavior mimicry[12], is an important yet underexplored principle in studying online community participation. However, homophily is a well-established principle in the context of social network analysis[13], and it has been shown to positively correlate with credibility of authors in online health communities[14]. Moreover, homophily expressed through unconscious mimicry was correlated with likelihood of liking the respondents[15]. Similarly in language, verbal mimicry of function words (i.e., content-free parts of speech) has been shown to positively correlate with liking the respondents[16] and positively functioning social dynamcs[17]. In online health communities, both function words and content words can serve as important cues for measuring homophily expressed in vocabulary usage. The function words are related to unconscious mimicry whereas the content words are related to similarity in health traits. Although homophily measured using all types of vocabulary usage—vocabulary similarity—could have effects on members' active participation, vocabulary similarity has not been studied with respect to active participation in online health communities.

Based upon previous literature, we expect community members to appreciate responses in which respondents use similar vocabulary. Thus, we hypothesized that individuals are likely to sustain active participation if they receive replies written with similar vocabulary. We examined this issue within the context of five online health communities from WebMD.com to address the following research questions:

(**RQ1**) What is the relationship between receiving replies written using a similar vocabulary and original posters' subsequent thread engagement?

(**RQ2**) What is the relationship between receiving replies written using a similar vocabulary in the early stage of joining the community and newcomers' sustained community participation?

(**RQ3**) What factors other than homophily in vocabulary usage correlate with active participation in online health communities?

METHODS

Our overarching objective is to understand participation in online health communities, in particular the relationship between vocabulary similarity of received replies and the member's future interaction in the community. We define **original post** as a post that starts a thread and **original poster** as the author of the original post. Similarly, we define **first reply** as the first post to reply to an original post and **respondent** as the author of the first reply. If the original poster or respondent uses multiple posts consecutively, we considered the accumulation of those posts as the original post or first reply, respectively. For example, original posters and respondents occasionally add comments in subsequent posts before any other member replies. Hence we included any supplementary posts as a part of the original post/first reply. We define **reengagement** as the behavior of original posters returning back to threads they started and having further conversation with the respondent (i.e., by posting a reply). Conversely, we defined **disengagement** as the behavior of original posters not posting a reply to that thread.

In our analysis, we restricted our focus to the first reply for two reasons. First, we wanted to pick the post with the highest chance of reaching the original poster. First replies appear for the longest time compared to other posts in the thread. Hence original posters have the longest time to read first replies. Second, systematically assessing who is responding to whom is difficult without analyzing the content of each post. For instance, the third person to post (the second replier) could be interacting with the respondent or the original poster. Those posts that are not replying back to the original post could skew the results; thus, we focused our analysis on first replies.

We reviewed common approaches from information retrieval that could be used to quantify **vocabulary similarity** that would represent homophily of vocabulary usage between original posters and respondents. We decided to use a vocabulary-based cosine similarity measurement without any feature reductions (e.g., removing common words) to quantify vocabulary similarity score. We chose to use cosine similarity because it is one of the most common and thoroughly studied measures[18]. One advantage of cosine similarity over other text similarity measures, such as Jaccard similarity, is that cosine similarity normalizes the text length during the comparison. Thus, long first replies would not necessarily be considered to have higher number of shared words. To determine the cosine similarity between original posts and first replies, we first represent each post as vector in N-dimensional space, where N is the number of unique terms across all posts and the value is the frequency with which terms occur in that post. Cosine similarity measures the cosine of the angle between two vectors representing the posts. The resulting similarity score ranges from zero to one. A score of zero indicates no shared terms between the two posts, whereas a score of one indicates all terms and the relative proportion of the terms used are exactly equal.

(RQ1) What is the relationship between receiving replies written using a similar vocabulary and the original posters' subsequent thread engagement?

To examine RQ1, we investigated whether original posters reengaged or disengaged in the threads given the vocabulary similarity score of first replies to original posts. We applied statistical tests (i.e., Pearson's Chi squared test (X^2) and Welch's t-tests (t) for two unpaired samples with unequal variances) to determine whether original posters who received replies with higher similar vocabulary scores reengaged more often. Thus, we compared the mean vocabulary similarity score among original posters who reengaged with the mean vocabulary similarity score among original posters who disengaged.

Next, we used logistic regression to predict the likelihood of original posters reengaging in their threads given vocabulary similarity score. Logistic regression is a statistical technique for predicting dichotomous outcome variables (i.e., engagement) given one or more predictor variables (i.e., vocabulary similarity scores). Logistic

regression limits the range of outcome variables from zero to one, satisfying assumptions for dichotomous outcome. Then, we tested the overall effects of vocabulary similarity score using the Wald test.

(RQ2) What is the relationship between receiving replies written using a similar vocabulary in the early stage of joining the community and the newcomers' sustained community participation?

We applied survival analysis to examine the relationship between newcomers receiving replies written using a similar vocabulary to their own posts in the early stage of joining the community and the newcomers' sustained participation in the community over time. To identify newcomers, we selected members who contributed at least one original post and received at least one first reply in their newcomer stage. The threshold for the newcomer stage was defined as up to three original posts. We chose this threshold because members with less than three posts were considered lurkers, who are not yet a regularly contributing member, in a prior study[10].

Survival analysis is a time duration analysis that models survival time until the failure event occurs. We define the survival object (i.e., "sustained participation") as the period of time in which members continue to participate in the online health community. Defining survival time with respect to online participation can be difficult because the failure event cannot be as clearly defined as in other fields, such as biological and medical sciences where survival analysis has been widely used. In the context of online health communities, members can always return to the community after years of absence as long as the community is active. We adopted a definition of a failure event from Wang et al.[8] to be a period of inactivity of three months without posting to the community. We considered members' first post (i.e., either original post or replying post in threads) as the starting point of their participation in the community and their last post as the end of their participation. However, if members posted within three months of the data collection date, we considered them right censored (i.e., member who did not experience the failure event) because they might still be actively participating in the community. We calculated the survival time as the days between members' first and last post.

(RQ3) What factors other than homophily in vocabulary usage are correlated with active participation in online health communities?

We selected a random sample of 100 original-first reply post pairs by selecting 10 pairs that reengaged and 10 pairs that disengaged in each of the five communities. We manually examined these 100 threads for other factors related to active participation. We drew on findings from previous studies to guide our content analysis. Previous studies have shown that types of social support[19] sought by original posters (e.g., informational or emotional), types of social support that original posters received[8], length of original posts and first replies[20,21], and rhetorical elements (i.e., asking questions[9]) are associated with participation.

In addition, we considered coverage of information—whether replies address all of the concerns expressed in original posts. In information retrieval, cosine similarity is used to measure the similarity of two documents with respect to their subject[18]. Because cosine similarity can calculate homophily of vocabulary and similarity of two documents—a proxy for coverage of information—we investigate a possible correlation between information coverage and active participation to have a deeper understanding of the effects of homophily in vocabulary.

We blindly examined the effect of these factors on future interactions with respondents. Furthermore, we explored the purpose of original posters' reengagement. The review of types of emotional support and the purpose of original posters' reengagement followed an open coding process[22], which is a method used to elicit unknown, emerging themes grounded in data. For informational support, we assessed whether the original posters were seeking information or not.

Data: Selection and Overall Characteristics

To meet our study aims, we restricted our analysis to five communities from WebMD.com for several reasons. First, we selected chronic disease-related communities. This criterion eliminated non-disease specific communities, such as parenting and *baby's first year* or smoking cessation, with a possible correlation between short-term health issues and dropout rates of WebMD forums. Second, we selected highly active communities that ranked within the top 20 WebMD forums in total number of threads. This eliminated communities that could have member dropouts due to the low-activity of the community. Third, we selected communities with two or more moderators who helped as respondents. This eliminated communities that could have member dropouts due to the low-level of moderating. Lastly, we selected communities with a sufficient number of posts from members, at least 50 first replies from both members and moderators. After applying these inclusion and exclusion criteria, five communities remained eligible for subsequent analysis: (1) ADHD, (2) Diabetes, (3) Heart Disease, (4) Pain Management, and (5) Sexual Health

communities (Table1). We excluded two types of original posts: (1) original posts without replies and (2) original posts started by moderators. We removed these types of original posts to focus our analysis on members and their participation behavior when receiving replies. We sought review by University of Washington Institutional Review Board (IRB) and the data was exempt from review.

Table 1. Characteristics of five WebMD communities studied

	ADHD	Diabetes	Heart Disease	Pain Management	Sexual Health
Dates data was collected	7/ 2005-6/2012	6/2007-5/2012	3/2008-5/2012	9/2007-6/2012	9/2007-1/2013
# threads analyzed	1,655	4,964	3,368	3,350	7,162
# members as original posters	1,484	2,459	2,817	2,752	5,766
# members as respondents	340	229	129	426	1,238

RESULTS

First reply distribution

First replies in our data set were posted within a day (mean of 21 hours), and 99% were posted within a week. Although replies posted later have an increased chance of not reaching the original posters, the reengagement rate for late replies that came after a week (9%) was comparable to the entire dataset's reengagement rate (17%). Thus, all first replies were included in the following analysis.

Results for (RQ1): What is the relationship between receiving replies written using a similar vocabulary and the original posters' subsequent thread engagement?

As shown in Table 2, we observed zero-inflated data distribution in which frequent zero vocabulary similarity scores were detected. Zero vocabulary similarity scores often resulted from the limited terms posted by respondents in first replies (e.g., *"Find another doctor"* or *"no comment"*). Original post and first reply pairs with zero vocabulary similarity scores had mean of 14 terms (Standard Deviation (SD)=29) in the first replies whereas the pairs with non-zero vocabulary similarity scores had significantly higher mean of 129 terms (SD=134; $t(800)=70.21$, $p<0.0001$) in the first replies. Because short generic replies are common in online communities, it was important that we keep the posts with zero similarity scores. We solved the high number of zero vocabulary similarity scores problem by fitting the data into a two-part model (i.e., zero-inflated continuous data model). We then analyzed original post and first reply pairs with zero and non-zero vocabulary similarity scores separately.

In the first part of the two-part model, we compared reengagement associated with vocabulary similarity score of zero versus non-zero. We compared the reengagement rate between the zero data set and the non-zero data set, which was significant ($X^2(1, N=20,499) = 15.27$, $p<0.0001$). Thus, having any vocabulary similarity score is associated with significantly higher reengagement.

Table 2. Proportions of original post-fist reply pairs with zero and above zero vocabulary similarity score by type of engagement by the respondent.

	# zero scores	# non-zero scores	Total # of pairs
Reengagement	51	4,330	4,381
Disengagement	336	15,782	16,118

In the second part of the model, we applied Welch's t-tests to the non-zero portion of data to compare reengagement and disengagement by the original poster. Overall, Welch's t-tests showed significantly higher vocabulary similarity score for reengagement compared to disengagement by the original poster in all communities except for Heart Disease (Table 3). Table 3 also shows the percentage of threads that original posters reengage later in the thread after multiple posters have posted. This data shows a full picture of original posters' engagement behaviors. We suspect that no difference was found in the Heart Disease community because of its overall higher rate of disengagement. In the Heart Disease community, original posters disengaged 81% of threads, which is 20% higher than the average disengagement rate of 60% in the other four communities.

Table 3. Comparison of mean similarity and by type of engagement, and percentage of reengagement by original posters, disengagement by original posters, and reengagement by original posters later in the thread after multiple posters have posted

	Mean (SD) vocabulary similarity for reengagement	Mean of vocabulary similarity for disengagement	Comparison of vocabulary similarity score: Reengagement vs. disengagement	% of reengaging	% of disengaging	% of reengaging later in thread
All five communities	0.38(0.15)	0.35(0.15)	t(6,637)=13.45 p<2.2e-16	17%	63%	20%
ADHD	0.42(0.16)	0.38(0.15)	t(370)=4.07 p=5.8e-05	15%	76%	9%
Diabetes	0.34(0.15)	0.32(0.15)	t(2,186)=5.58 p =2.687e-08	19%	52%	30%
Heart Disease	0.38(0.13)	0.37(0.12)	t(680)=1.58 p =0.11	14%	81%	5%
Pain Management	0.43(0.15)	0.37(0.15)	t(1,242)=9.61 p < 2.2e-16	19%	63%	18%
Sexual Health	0.39(0.16)	0.34(0.15)	t(2,351)=10.80 p < 2.2e-16	17%	62%	20%

Next, we used logistic regression with one predicator variable—vocabulary similarity score—to predict the likelihood of original posters reengaging in their threads. Figure 1 shows the plot of the predicted probability with 95% confidence intervals of reengagement given the vocabulary similarity score between original post and first reply. This regression model was significant ($X^2(1)=91.43$, p=1.55e-43) with odds ratio of 4.70. Using the vocabulary similarity score, we predicted future participation with 79% accuracy in a 10-fold cross validation. The Wald test indicated that for a one unit increase in vocabulary similarity score, the odds of original posters reengaging increased by a factor of 4.7 ($X^2(1)=188.60$, p=6.43e-43).

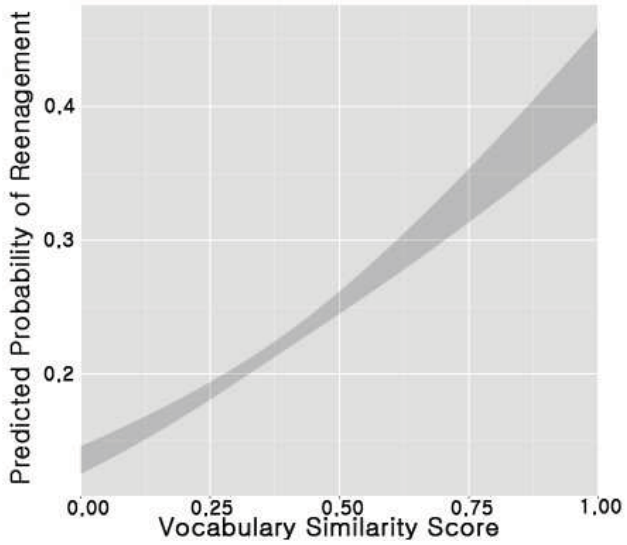

Figure 1. Predicted probabilities of reengagement graph with 95% confidence intervals.

Results for (RQ2): What is the relationship between receiving replies written using a similar vocabulary in the early stage of joining the community and the newcomers' sustained community participation?

We applied survival analysis to test the effect of receiving replies written using a similar vocabulary on members' sustained participation in the community. We partitioned members into three equally sized groups corresponding to members exposed to replies with a "High," "Medium," or "Low" vocabulary similarity score. For members with more than one original and corresponding first reply, we took the average vocabulary similarity score among the first three original and corresponding first replies. Low vocabulary similarity scores ranged from 0 to 0.28, Medium scores ranged from greater than 0.28 to 0.41; and high scores ranged from greater than 0.41 to 0.83, which was the highest vocabulary similarity score in our dataset. Examples of high and low replies are shown in Results section for RQ3.

Figure 2 illustrates the effect of receiving replies written using a similar vocabulary on members' sustained active participation. Members in the High group were most likely to stay active in the community, followed by members in the Medium group, followed by members in the Low group as least likely to stay active. These differences were sustained between the high and low groups for at least 300 days.

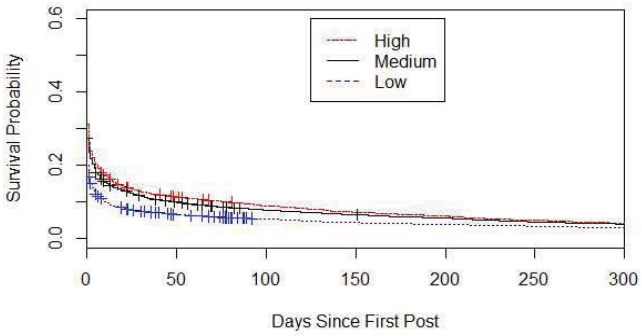

Figure 2. Survival curves for members exposed to high, medium, and low levels of vocabulary similarity in replies

Results of two survival models are shown in Table 4. Model 1 reports the effects of the covariates. For instance, the hazard ratio of 0.75 for the total number of original posts indicates that those who initiate threads one standard deviation more have a 34% (i.e., (1/0.75) – 100%) higher survival rate. Similarly, Model 2 shows that members who received replies with a vocabulary similarity score of one standard deviation higher have a 5% (i.e., (1/0.95) – 100%) higher survival rate when controlling for covariates.

The hazard ratio indicates the odds of members dropping out of the community (encountering the failure event). We also considered a number of covariates and their relationship to sustained participation in two survival models. We selected covariates representative of each member's intrinsic characteristics (e.g., sociable) as well as the amount of participation in the community. These variables include the total number of posts, total number of first replies provided, total number of first replies received, and total number of original threads. We normalized variables (i.e., (observation – mean)/standard deviation) to show predicted change in odds for a unit increase in the predictor.

Table 4. Survival analysis showing influence of covariates in two models

	Model 1		Model 2	
Covariates	**Hazard Ratio**	**Standard Error**	**Hazard Ratio**	**Standard Error**
Total number of posts	0.91**	0.034	0.92**	0.034
Total number of first replies provided	0.92**	0.032	0.92**	0.031
Total number of first replies received	0.83*	0.091	0.84	0.091
Total number of original threads	0.75**	0.095	0.74**	0.094
Vocabulary similarity scores			0.95***	0.008

*****: p<0.001, **: p<0.01, *: p<0.05**

Results for (RQ3): What factors other than homophily in vocabulary usage are correlated with active participation in online health communities?

Without any knowledge of their vocabulary similarity scores or reengagement status, we manually categorized original post and first reply pairs into three groups: high, medium, and low coverage groups. We categorized pairs with first replies that addressed all of the concerns expressed in original posts as **high coverage**, first replies that addressed some concerns as **medium coverage**; and first replies that did not address any concerns as **low coverage**. We then examined how well the vocabulary similarity measures performed compared to manual categorization. High coverage group compared to low coverage show significantly higher vocabulary similarity scores ($t(19)=2.58$, $p=0.02$) (Table 5). However, the difference between high coverage group and medium coverage group ($t(20)=0.34$, $p=0.74$) as well as the difference

Table 5. A comparison among subjective and vocabulary similarity scores

	Mean of high coverage (SD)	Mean of medium coverage (SD)	Mean of low coverage (SD)
Vocabulary similarity scores	0.40 (0.14)	0.39 (0.18)	0.29 (0.16)

between medium coverage group and low coverage group (t(29)=1.63, p=0.11) were not significant.

In the 50 reengaging pairs, we found that 86% provided high coverage, 8% provided medium coverage, and 6% provided low coverage. In contrast, of the 50 disengaging pairs, 52% provided high coverage, 24% provided medium coverage, and 24% provided low coverage. Comparison of high, medium, and low information coverage showed reengaging had a significantly higher proportion of high coverage replies (X^2(1, N=100)=11.97, p=0.0005). As well, disengaging had a significantly higher (X^2(1, N=100)=5.02, p=0.03) proportion of low coverage replies. However, medium coverage did not differ significantly (X^2(1, N=100)=3.65, p=0.06).

Similarly, when we compared pairs associated with reengagement and with disengagement, we found a noticeable difference in the length of the posts measured by number of words. The pairs associated with reengagement used more words in both original posts and first replies. On average in the reengaging pairs, original posters used 198 words (SD=152) while respondents used 181 words (SD=150). In contrast, in the disengaging pairs, original posters used 122 words (SD=116) while respondents used 137 words (SD=142). The difference between reengaging and disengaging pairs was significant in original posts (t(92)=2.78, p=0.006) but not significant in first replies (t(98)=1.48, p=0.14).

We also found differences in emotional support: whether respondents indicated an aspect of empathizing with or helping original posters. We found two themes of emotional support: **acknowledgement** of members' experience of difficulty (e.g., *"I know exactly how you feel* [...] *I'm in the same boat"*) and **encouragement** for the current situation (e.g., *"You've got a great attitude, and will do well"*). In reengaging pairs, 30% of respondents acknowledged the difficulty of the original poster's situation and 34% of respondents encouraged original posters. Conversely, in disengaging pairs, 18% and 28% of respondents acknowledged their difficulty and provided encouragement, respectively. However, differences in these propositions were not significant (Acknowledgement: X^2(1, N=100)=1.37, p=0.24; Encouragement: X^2(1, N=100)=0.19, p=0.67).

The overall exchange of emotional support was less frequent than informational support; this fits well with what original posters were seeking. Overall, 89% of the original posters asked for new information, while only 24% of the original posters showed any signs of requesting emotional support (e.g., *"can't take more"* or *"desperate!"*). Exchanges of emotional and informational support were not mutually exclusive as some original posters sought both.

We also examined the effects of respondents asking questions to original posters in first replies. In reengagement, 32% of respondents asked a question in their reply, while only 20% of respondents asked a question in disengagement, however, difference in these proportions were not significant (X^2(1, N=100)=1.30, p=0.25). Similarly difference in the proportions of providing informational support was not significant (X^2(1, N=100)=0.07, p=0.79) between reengagement and disengagement. In reengagement, 84% provided informational support, while 80% provided informational support in disengagement.

In our qualitative analysis, we identified four themes for the purpose of original posters' reengagement: (1) providing more information on their situation (82%), (2) thanking the original posters (62%), (3) asking more questions (20%), and (4) getting defensive (8%). These behaviors were not mutually exclusive.

In our sample, reengaging and disengaging pairs varied in the degree of information coverage. First replies that covered more aspects of the original post received higher vocabulary similarity scores than those that covered fewer aspects. The following is an example of a reengaging pair that received a relatively high vocabulary similarity score of 0.66.

Original poster_A: *"I'd much rather ask others that have a condition I may have... I've been pretty hyper all my life […] and have been bouncing around the idea in my head that I may have some sort of ADHD. […]I've noticed I can focus more when I am very tired. […]"*

Respondent_to_A: *"You are describing Adult ADHD. You should make an appointment with a psychiatrist who was experience with adult ADHD. […] My mind is usually all over the place bouncing from topic to topic […] Things aren't like that now. […] Go. Get tested. […] If you do have ADHD, he may or may not prescribe medication. […]You have nothing to loose as long as your honest with your doctors. Good luck. […]"*

Conversely, the following is an example of a disengaging pair that received relatively low vocabulary similarity score of 0.14.

Original poster_B: *"been having muscle ache between upper left chest (near armpit) through shoulder […] Did I* strain a muscle/group of muscles, and what's best - *heat, NSAIDS,chiropratic?"*

Respondent_to_B: *"Hello. No way for any of us on an internet message board to know what is causing your symptoms. You will need to see your doctor for an evaluation and treatment plan […]"*

In both examples, the respondents advocate seeking professional help. However, only in the first example, did respondent_to_A provide their experience and perspective. During this process, respondent_to_A covered a number of issues raised by original poster_A while using more shared vocabulary. This conversational pair resulted in a relatively high vocabulary similarity score and elicited original poster_A to reengage in further conversation. In the second example, respondent_to_B does not address the concerns original poster_B raised. This resulted in using less shared vocabulary with original poster_B and a relatively low vocabulary similarity score.

As mentioned earlier, other factors also correlate with original posters' engagement. Respondent_to_A acknowledges the difficulty and encourages original poster_A while using more words (521 words) than respondent_to_B (34 words). Conversely respondent_to_B is succinct and does not provide any emotional support while neither respondent asked questions. Although a combination of many factors can influence engagement of the original poster, in our manual analysis receiving replies with more shared vocabulary appears to be associated with reengagement, which supports our quantitative analysis.

DISCUSSION AND FUTURE WORK

Prior research shows that sustaining active participation presents a prominent challenge for online communities[6–10]. In this paper, we showed the importance of homophily expressed through shared vocabulary associated with members' ongoing engagement in online health communities. Members who received replies that contained more shared vocabulary with their own posts tended to continue their conversations with respondents.

Additionally, we created prediction models that estimate the likelihood that original posters will reengage with respondents. Although many factors can contribute to members disengaging from conversation, our logistic regressions showed that vocabulary similarity predicts future participation with the respondents. Similar prediction models could be one solution for the challenge of sustaining active participation. For instance, our prediction model could allow moderators to identify members who are most likely to disengage. This added knowledge could enable moderators to provide replies that encourage reengagement. Moreover, we discovered that receiving replies written in a similar vocabulary in the early stages of joining the community predicts long-term active participation within the community. One solution for sustaining newcomers' participation in the community is to encourage community members to abide by a set of guidelines, which reflect ideal community member interactions with newcomers. Furthermore, measuring vocabulary similarity can be a basis for automatically alerting members when they deviate from posting replies that encourage reengagement. In contrast with targeting specific members to encourage reengagement, the environment of the community as a whole could be improved by filtering spam or abusive content through comparing terms with the common vocabulary of the community.

In our manual analysis, we found that a combination of many different factors could influence original posters' engagement. Other factors, such as exposure to higher degree of information coverage and higher word counts by original poster[20,21] were associated with reengagement in our data. All other factors had higher occurrences in reengagement, but they are not found significant. We suspect this is due to small sample size of 100 in our manual analysis. Investigating which factors had the biggest impact on original posters' engagement is an important question for future work. However, measuring vocabulary similarity of first reply to the original post is a relatively easy and robust technique that seems to measure an important marker for member reengagement.

Furthermore, the vocabulary similarity of the first reply might not be the sole factor of member engagement with respondents. For instance, other components of life can influence online health community participation. Original posters could have gained the needed information through other sources and did not check back with the community. A serious medical crisis could prevent original posters from checking back with the community too, for example. Still, the consistent statistical results from examining the relationship between vocabulary similarity and participation show that homophily of vocabulary provides an important marker for member reengagement.

Our survival analysis examines members' early stages of joining the community and uses a threshold of the first three replying posts that members received based on previous literature on lurking[10]. Understanding the transition point of members' participation can provide a deeper understanding of how to sustain active participation in online

communities. Also, our analysis does not answer whether inactive members remained lurkers or completely dropped from the community. An analysis of these two different types of inactive members could further extend our findings.

Although overall our analyses showed consistent results in a diverse group of online health communities, we acknowledge that our large sample size could have inflated the significance levels or otherwise skewed results and raises questions to the practical significance of our findings. Also, how much vocabulary similarity is needed for a meaningful increase in reengagement remains an open question. Further investigation, such as surveys or interviews, analyzing members' satisfaction and understandability of health information correlated to vocabulary similarity could further explore the significance of our findings.

In future work, we plan to investigate the correlation between actual users' perceived qualities of replies and participation to further examine the challenge of sustaining participation in online health communities. Understanding these relationships could provide a more complete view of how to sustain participation in online health communities. The significance of this study goes beyond predicting members' behavior in online health communities. For instance, our findings could generalize in non-health online communities and our automatic approach to analyze computer-mediated communication (CMC) could be applied to other CMC studies. Furthermore, our study showed a potential method to elucidate the process of forming social bonds through CMC in online communities.

CONCLUSION

We provide new insights regarding sustaining online health community participation through systematic analyses of five WebMD online health communities. Our findings suggest that homophily—the vocabulary similarity between members' posts—plays a crucial role in sustained engagement in online health community. We provide new insights into how vocabulary similarity affects active participation in online communities. Furthermore, vocabulary similarity calculated with cosine similarity shows promising results in measuring the coverage of information in replies. Based on these insights, moderators, online community creators, and online community participants could tailor replies to encourage sustained, active participation by members. Findings from this study can improve member experience in difficult situations when online health communities provide essential support.

ACKNOWLEDGEMENTS

This work was funded by NSF SHB 1117187 and NIH-NLM # 5T15LM007442-10 BHI Training Program.

REFERENCES

1. Fox S, Duggan M. Health online 2013: 35% of U.S. adults have gone online to figure out a medical condition; of these, half followed up with a visit to a medical professional. Health (Irvine Calif) [Internet]. 2013;1–55. Available from: http://www.webcitation.org/6Y5K9a3JH
2. Van Uden-Kraan CF, Drossaert CHC, Taal E, Shaw BR, Seydel ER, van de Laar M a FJ. Empowering processes and outcomes of participation in online support groups for patients with breast cancer, arthritis, or fibromyalgia. Qual Health Res [Internet]. 2008 Mar [cited 2014 Jan 12];18(3):405–17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18235163
3. Van Uden-Kraan CF, Drossaert CHC, Taal E, Seydel ER, van de Laar M a FJ. Participation in online patient support groups endorses patients' empowerment. Patient Educ Couns [Internet]. 2009 Jan [cited 2014 Jan 12];74(1):61–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18778909
4. Setoyama Y, Yamazaki Y, Namayama K. Benefits of peer support in online Japanese breast cancer communities: differences between lurkers and posters. J Med Internet Res [Internet]. 2011 Jan [cited 2014 Jan 12];13(4):e122. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3278108&tool=pmcentrez&rendertype=abstract
5. Bartlett YK, Coulson NS. An investigation into the empowerment effects of using online support groups and how this affects health professional/patient communication. Patient Educ Couns [Internet]. Elsevier Ireland Ltd; 2011 Apr [cited 2014 Jan 12];83(1):113–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20599338
6. Millen D, Patterson J. Stimulating social engagement in a community network. Proceedings of the 2002 ACM conference on Computer supported cooperative work [Internet]. ACM; 2002 [cited 2014 Jan 13]. p. 306–13. Available from: http://dl.acm.org/citation.cfm?id=587121
7. Joyce E, Kraut RE. Predicting Continued Participation in Newsgroups. J Comput Commun [Internet]. 2006 Apr [cited 2014 Jan 1];11(3):723–47. Available from: http://doi.wiley.com/10.1111/j.1083-6101.2006.00033.x

8. Wang Y, Kraut R, Levine JM. To Stay or Leave ? The Relationship of Emotional and Informational Support to Commitment in Online Health Support Groups. Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. ACM; 2012.
9. Arguello J, Butler B, Joyce E, Kraut R, Ling KS, Rose C, et al. Talk to me: foundations for successful individual-group interactions in online communities. Proceedings of the SIGCHI conference on Human Factors in computing systems [Internet]. ACM; 2006 [cited 2014 Jan 1]. p. 959–68. Available from: http://dl.acm.org/citation.cfm?id=1124916
10. Nonnecke B, Preece J. Lurker demographics: Counting the silent. Proceedings of the SIGCHI conference on Human factors in computing systems [Internet]. ACM; 2000 [cited 2014 Apr 27]. p. 73–80. Available from: http://dl.acm.org/citation.cfm?id=332409
11. Roberts TL. Are newsgroups virtual communities? Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '98 [Internet]. ACM Press/Addison-Wesley Publishing Co.; 1998. p. 360–7. Available from: http://portal.acm.org/citation.cfm?doid=274644.274694
12. Granitz N a., Koernig SK, Harich KR. Now It's Personal: Antecedents and Outcomes of Rapport Between Business Faculty and Their Students. J Mark Educ. 2008;31:52–65.
13. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: Homophily in social networks. Annu Rev Sociol. 2001;27:415–44.
14. Wang Z, Walther JB, Pingree S, Hawkins RP. Health information, credibility, homophily, and influence via the Internet: Web sites versus discussion groups. Health Commun. 2008;23(4):358–68.
15. Chartrand T, Bargh J. The chameleon effect: The perception–behavior link and social interaction. J Pers Soc Psychol [Internet]. 1999 [cited 2014 Jul 16];76(6):893. Available from: http://psycnet.apa.org/journals/psp/76/6/893/
16. Ireland ME, Slatcher RB, Eastwick PW, Scissors LE, Finkel EJ, Pennebaker JW. Language style matching predicts relationship initiation and stability. Psychol Sci [Internet]. 2011 Jan [cited 2014 Jul 16];22(1):39–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21149854
17. Gonzales a. L, Hancock JT, Pennebaker JW. Language Style Matching as a Predictor of Social Dynamics in Small Groups. Communic Res [Internet]. 2009 Nov 4 [cited 2014 Jul 21];37(1):3–19. Available from: http://crx.sagepub.com/cgi/doi/10.1177/0093650209351468
18. Singhal A. Modern information retrieval: A brief overview. IEEE Data Eng Bull [Internet]. 2001 [cited 2014 Jul 16];1–9. Available from: http://act.buaa.edu.cn/hsun/IR2013/ref/mir.pdf
19. Cutrona CE, Suhr JA. Social support communication in the context of marriage: An analysis of couples' supportive interactions. 1994.
20. Adamic LA, Zhang J, Bakshy E, Ackerman MS. Knowledge sharing and yahoo answers: everyone knows something. Proceedings of the 17th international conference on World Wide Web [Internet]. ACM; 2008 [cited 2014 Feb 28]. p. 665–74. Available from: http://dl.acm.org/citation.cfm?id=1367587
21. Agichtein E, Castillo C, Donato D, Gionis A, Mishne G. Finding high-quality content in social media. Proceedings of the international conference on Web search and web data mining - WSDM '08 [Internet]. New York, New York, USA: ACM Press; 2008. p. 183. Available from: http://portal.acm.org/citation.cfm?doid=1341531.1341557
22. Strauss AL, Corbin JM. Basics of qualitative research. Newbury Park, CA: Sage Publications; 1990.