Text Analysis as a Tool for Analyzing Conversation in Online Support Groups

Adam D. I. Kramer, Susan R. Fussell, Leslie D. Setlock
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
[adk, sfussell, lsetlo]@cs.cmu.edu

ABSTRACT
In this paper we describe a software tool that allows investigators to make comparisons across different online forums and media by analyzing word counts in user-specified categories. Using a large sample of messages from a bipolar support chatroom, we demonstrate how this tool can be used to characterize the nature of the discourse and compare it to other media, to analyze relationships among different word categories, and to characterize changes in visitors' discourse over time. Future plans for adding functionality to the software and using external data for additional validation are also discussed.

Author Keywords
Online communities, computer-mediated communication, conversational analysis, empirical studies, computer-supported cooperative work

ACM Classification Keywords
H5.3 Group and Organizational Interfaces: *Computer-supported cooperative work, synchronous interaction, evaluation/methodology*

INTRODUCTION

Online support is a growing forum for self-help among sufferers from chronic mental and physical illnesses. By providing 24-hour access to others with the same problem, online support groups are hypothesized to reduce social isolation, improve coping skills, and improve patients' knowledge about their condition. To date, however, these hypotheses have been neither confirmed nor disconfirmed. On one hand, studies of specialized online support communities have suggested that the groups are valuable for specific populations [e.g., 1, 3, 6, 11]. On the other hand, larger surveys of Internet users have found negative psychological effects of extensive online participation [8].

One reason for the unclear pattern of results is a lack of adequate measures. Few previous studies have provided measures of the value of online interaction that can be applied across different medical conditions and different communications media. Without such measures, however, it is difficult to determine whether online support should be recommended by therapists, physicians, and other professionals as an adjunct to traditional treatments.

The difficulty in establishing concrete measures is directly related to one of the attractions of online support communities. Online groups are anonymous, making it difficult for researchers to contact participants for additional information (e.g., to complete surveys or collect objective behavioral measures). This information is required to determine, for example, whether involvement in a support group increases adherence to prescribed treatment regimens.

In this paper we describe an approach to evaluating the effectiveness of online support based on analysis of what people talk about in a chatroom. In contrast to previous work on structural features of online conversations [e.g., 2, 7], we focus on the content of online conversation. We describe a software tool written by the first author that analyzes chatroom discourse in terms of word categories, similar to Pennebaker and Francis' Linguistic Inquiry and Word Count (LIWC) tool [9]. We explain how we generated word categories for our analyses and provide some illustrative ways in which we've used this tool to characterize discourse in a bipolar disorder support chatroom. We conclude with directions for future research.

METHOD

Message Sample

In this analysis, we analyze communication in a long-running bipolar disorder chatroom. The chatroom goals were to inform newly diagnosed individuals about treatment options, provide information to friends and family members, and supply ongoing peer-to-peer support.

CHI 2004, April 24–29, 2004, Vienna, Austria.
ACM 1-58113-703-6/04/0004.

We examined a sample of 80,136 individual chat messages recorded on 289 different days between July 7, 2002 and October 11, 2003. Messages were generated by a total of 267 different visitors, who visited the room on between 2 and 109 of the days during the sampled period. Each participant contributed an average of 206.42 words per visit (SD = 272.60). For the word count analyses, we computationally analyzed all logs, resulting in 2827 person-by-day entries.

Word Count Tool

Our word count tool works similarly to the LIWC [9] but contains enhancements to standardize the processing of logs from the mIRC Internet Relay Chat (IRC) client. Processing occurs in two phases, both implemented in PERL.

In the *preprocessing phase*, system messages and other extraneous messages are removed from the IRC logs. A combination of automated and manual procedures is used to identify nicknames referring to the same individual. For example, if someone with the nickname "Murphy" changes his nickname to "Murphy-brb," messages under the second nickname would be attributed to the first, canonical nickname. IP address matching is also used to identify the same individual under different names. The preprocessing phase generates files in which each line contains time and date stamps, a canonical nickname, and one chat utterance.

In the *analysis phase*, each line of the chat log is analyzed separately. Each word in each line is compared against each word in user-specified dictionaries. Dictionary and category files were similar to the LIWC format, in which a set of word categories are given numerical designations and then each word appears with the appropriate numbers after it. We provide additional support for full PERL regular expressions and word sequences. Word counts are collapsed across day, and across nicknames within individuals. The result is a data file comprised of a line for each canonical nickname for each day, a word count for each category, and a count of total words used that day, per nickname.

Word Categories

Categories of words used in this analysis were drawn from two sources. First, we developed a core set of terms related to bipolar disorder. A previous study of the same bipolar chatroom [5] found that much of the conversation focused on treatment methods (e.g., medications, therapy) and ways of coping with bipolar disorder in everyday life (e.g., family relations, working). To generate word categories, we examined a subset of the chat logs for words and phrases characteristic of discussion of bipolar disorder and grouped these words into subcategories. Then, two other individuals with expertise in bipolar disorder checked the words and suggested additional terms where necessary. The resultant nine categories are shown in Table 1. We also created a *bipolar* meta-category containing all terms in the nine categories.

Word Category	Examples
Anger	Anger, hate, rage, yelling
Anxiety	Disturbed, anxious, freakout, scared
Depressed	Low, suicide, cried, blues, dysphoria
Disability	Medicaid, ssdi
Disorders	Bipolar, psychosis, nervous breakdown
Illness	Ill, sick
Mania	Hyper, euphoric, out of control
Medication	Antidepressant, stabilizer, side effect, dose
Treatment	Doctor, nuthouse, untreated, psychiatric

Table 1. Categories of words associated with Bipolar disorder with examples of terms in each category.

We also examined a subset of the word categories used by Pennebaker and Francis in the LIWC [9], which has been used in previous studies of face-to-face conversation, written essays and books, and online discussions [e.g., 10, 12, 13]. (See Table 2.) Note that unlike our bipolar-related categories, some LIWC categories are not mutually exclusive (i.e., a word may appear in multiple categories).

Word Category	Examples
Pronouns	
I	I, my, me
We	We, our, us
You	You, you'll
Other	He, she, them
Negative Emotions	Worthless, hate, tense
Positive Emotions	Joy, love, good
Cognitive Processes	Know, think, consider
Social Processes	Converse, share, friends

Table 2. Selected categories from Pennebaker and Francis' LIWC used in the current study.

RESULTS

We consider the results in three parts. First, we use our word count system to characterize the nature of the conversations and compare them to discourse in other genres. Second, we examine relationships among word categories to understand how topics covaried in the sample. Finally, we demonstrate how our tools can be used to analyze changes in individual's communication over time.

Characterizing the nature of discourse

Because our tool can use both custom and established dictionary files, it is possible to characterize the nature of online conversations and compare them to other genres and technologies. In Table 3, for example, we show means for selected word categories in our sample, along with previously reported means from the domains of emotion writing and casual conversation [9]. These comparisons are helpful in understanding the nature of online discourse. For

example, we see that conversation in the bipolar chatroom is no more negative, emotionally, than other genres. It also appears to be less self-focused and more focused on social processes. Interestingly, terms specifically related to bipolar disorder and its treatment comprised a very small percentage of the overall discourse.

	Bipolar Chat[a]	Emotion Writing[b]	Talking[b]
Total Pronouns	11.51	17.2	15.8
I	5.23	10.6	5.6
We	.04	.8	1.0
You	2.11	.4	4.0
Other	1.38	3.3	2.5
Negative Emotions	1.66	2.6	1.3
Positive Emotions	3.45	2.7	2.7
Cognitive Processes	4.79	7.8	7.3
Social Processes	13.39	9.5	10.9
Bipolar Disorder	1.40	N/A	N/A

Table 3. Comparison of percentages of words in different categories from our study (a) and from previous research on emotion writing and casual conversation [9] (b).

Analyzing relationships among word categories

When word categories are disjoint, we can also examine relationships among different types of words. Although many LIWC categories [9] are not disjoint, our bipolar-related categories do not overlap with one another or with the LIWC pronoun categories. They do overlap to a small extent with the negative and positive emotion categories.

We examined correlations among proportions of bipolar words as a set and proportions of *I, we, negative emotions,* and *positive emotions* in the LIWC word set (Table 4). Consistent with previous work [e.g., 12], pronoun use was associated with emotional valence. Proportion of *I* words was negatively correlated with proportion of *you* words, and positively correlated with proportion of negative emotional state words. Proportion of *you* words was positively correlated with positive emotion-state words. Proportion of bipolar-related words was positively correlated with *I* words and negative emotion words, and negatively correlated with positive emotion words.

	Bipolar	I	You	Negative Emotion	Positive Emotion
Bipolar	1				
I	.12*	1			
You	-.02	-.05*	1		
Neg Emo	.16*	.13*	.06	1	
Pos Emo.	-.07*	-.03	.27*	-,05	1

Table 4. Correlations between bipolar-related words and selected LIWC categories. NOTE: *p < .01; N = 2822.

Characterizing change over time

Perhaps the most powerful application of word count analysis is in analyzing changes over time as visitors return to the chatroom. To perform this analysis, we assigned a visit day to each person's daily counts, depending upon the number of times they had previously visited the room. Persons who had visited the room repeatedly prior to the sampled period were excluded from this analysis.

Overall, words per day increased slightly but significantly with visit day ($r = .07$, $p < .0005$). We thus controlled for total number of words per person, per day, in our analyses of changes in word category usage. In addition, since the distribution of visits was skewed to the right, we used a logarithmic transformation of visit day in the analyses.

To illustrate the value of this approach, we examined predictors of the number of bipolar-related words in participants' messages. In Model 1 in Table 5, we show the results of regressing the count of bipolar words on log-days and total words spoken that day. Holding total words spoken per day constant, log of visits to the chatroom was associated with significantly fewer bipolar words used that day ($\beta = -.72$, $t = -4.67$, $p < .0005$).

Model 1	B	Std. Err	β	t
(Constant)	.74	.19		3.99**
Total Words	.02	.00	.76	62.17**
Visit Day (Log 10)	-.72	.15	-.06	-4.67**
Model 2				
(Constant)	.84	.18		4.70**
Total Words	.01	.00	.49	10.53**
Visit Day (Log 10)	-.59	.15	-.05	-4.00**
I	.06	.01	.21	6.20**
You	.05	.02	.06	3.09*
Negative Emotion	.25	.02	.27	10.45**
Positive Emotion	-.15	.02	-.22	-8.71**

Table 5. Regressions predicting number of bipolar-related words per user per day. Model 1 uses total words and log of visit day as predictors. Model 2 adds numbers of *I* words, *you* words, negative emotion terms, and positive emotion terms. NOTE: *p < .005; ** p < .001.

In Model 2, we add the *I* and *you* pronoun categories and the negative and positive emotion-state categories from the LIWC to our list of predictors. This model explains significantly more of the variance ($R^2 = .58$ vs. $.62$, $p < .0005$). All of these were significant predictors of the number of bipolar words used, and total words and log of visit day remained significant.

CONCLUSIONS AND FUTURE DIRECTIONS

The results demonstrate the utility of a word count analysis that takes into consideration individual-level variance for analyzing online discourse. We found that word analysis is helpful for characterizing the nature of discourse in the

sampled bipolar support chatroom and for comparing its content to communication in other genres. The same approach can be used to perform finer-grained analyses of online discourse. For example, we can compare bipolar chatrooms with different policies about who can participate and what can be discussed, examine similarities and differences across support chatrooms for different mental and physical disorders, and examine the effects of media (e.g., email distribution lists, online bulletin boards, chatrooms) on online support conversations.

We also found that the development of mutually exclusive word categories allowed us to further characterize the relationships among different categories of words in the chatroom. We plan to extend this analysis by incorporating conditional analysis into our software tool. Conditional analysis will help us disambiguate different uses of the same term, such as when *dark* or *drowning* are used as metaphorical descriptions of emotional states [cf. 4].

Finally, we have demonstrated the value of analyzing changes in individual-level word use over repeated visits to a chatroom. These changes are hypothesized to reflect changes to visitors' adjustment to their condition. In the next phase of this research, we will be testing this hypothesis by having a group of chatroom volunteers complete questionnaires about their treatment, emotional state, family relations, and other pertinent attitudes and behaviors over a six-month period while we continue to log their chatroom activities. Our goal is to further validate the relationships between discourse markers and client mental and physical health we have uncovered.

With regard to our word counting software, we plan to improve the computational efficiency and robustness of the text analysis process. We are also working to improve the word categorization system and to develop new word categories suitable for analyzing online support for other mental and physical conditions. Our long-term goals are to provide measurement techniques that can be used across a full range of mental and physical disorders and a full range of electronic discourse types and to identify those features of online support forums that lead to maximal improvement in mental and physical health.

ACKNOWLEDGMENTS

We thank Pamela Jackson for her help in obtaining the chatroom data and creating word categories. This research was supported by the Robert Wood Johnson Foundation e-Health Technologies Initiative grant #49139. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Robert Wood Johnson Foundation.

REFERENCES

1. Brennan, P. F., & Ripich, S. (1994). Use of a home-care computer network by persons with AIDS. *International Journal of Technology Assessment in Health Care, 10,* 258-272.
2. Cherny, L. (1999). *Conversation and community: Chat in a virtual world*. Stanford, CA: CSLI.
3. Epstein, Y. M., Rosenberg, H. S., Grant, T., & Hemenway, N. (2002). Use of the internet as the only outlet for talking about infertility. *Fertility and Sterility, 78,* 507-514.
4. Fussell, S. R., & Moss, M. M. (1998). Figurative language in descriptions of emotional states. In S. R. Fussell & R. J. Kreuz (Eds.), *Social and cognitive approaches to interpersonal communication*. Mahwah, NJ: Erlbaum.
5. Fussell, S. R., & Setlock, L. D. (under review). Informal communication in an online volunteer community: Implications for supporting virtual relationships.
6. Galegher, J., Sproull, L., & Kiesler, S. (1998). Legitimacy, authority, and community in electronic support groups. *Written Communication, 15,* 493-530.
7. Herring, S. C. (2001). Computer-mediated discourse. In D. Tannen, D. Schiffrin, & H. Hamilton, *Handbook of Discourse Analysis* (pp. 612-634). Oxford: Blackwell.
8. Kraut, R. E., Patterson, M., Lundmark, V., Kiesler, S., Mukhopadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist, 53*, 1017-1032.
9. Pennebaker, J.W., & Francis, M.E. (2001). *Linguistic Inquiry and Word Count: LIWC* [software program for text analysis]. Lawrence Erlbaum Associates.
10. Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology.*
11. Person, B. (2002). Online support for caregivers of people with a mental illness. *Psychiatric Rehabilitation Journal*, *26,* 70-77
12. Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and non-suicidal poets. *Psychosomatic Medicine, 63,* 517-522
13. Stone, L. D., & Pennebaker, J. W. (2002). Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana. *Basic and Applied Social Psychology, 24,* 173-183.