Analysis of Anxious Word Usage on Online Health Forums

Nicolas Rey-Villamizar
Dept. of Computer Science
University of Houston
Houston, TX 77204-3010
nrey@uh.edu

Prasha Shrestha
Dept. of Computer Science
University of Houston
Houston, TX 77204-3010
pshrestha3@uh.edu

Farig Sadeque
School of Information
University of Arizona
Tucson, AZ 85721
farig@email.arizona.edu

Steven Bethard
School of Information
University of Arizona
Tucson, AZ 85721
bethard@email.arizona.edu

Ted Pedersen
Dept. of Computer Science
University of Minnesota, Duluth
Duluth, MN 55812-3036
tpederse@d.umn.edu

Arjun Mukherjee, Thamar Solorio
Dept. of Computer Science
University of Houston
Houston, TX 77204-3010
arjun@cs.uh.edu, solorio@cs.uh.edu

Abstract

Online health communities and support groups are a valuable source of information for users suffering from a physical or mental illness. Users turn to these forums for moral support or advice on specific conditions, symptoms, or side effects of medications. This paper describes and studies the linguistic patterns of a community of support forum users over time focused on the used of anxious related words. We introduce a methodology to identify groups of individuals exhibiting linguistic patterns associated with anxiety and the correlations between this linguistic pattern and other word usage. We find some evidence that participation in these groups does yield positive effects on their users by reducing the frequency of anxious related word used over time.

1 Introduction

How people behave within a given community is an important question, especially in the context of health support. The advancement of technology has complemented the classic in-person health support forums to a growing and vibrant online community. Social media research has indicated that individuals psychological states and social support status relating to health and well-being may be deduced via analysis of language and conversational patterns (Tamersoy et al., 2015). In the offline world, some psychological studies of people's behavior have shown correlation of different sociological dimensions such as sadness and anger over the time course of a breakup process (Sbarra, 2006). We want to study these kinds of correlations in online support forums.

In this paper we focus our attention on the analysis of the users who participated in the DailyStrength forums[1], and we propose a methodology to study the users' behaviors by analyzing the linguistic characteristics of their posts. Researchers have shown that a large and increasing number of people are going online for medical information and advice (Fox and Duggan, 2013). We focus our study on the usage of words related to anxiety. This is an important area of interest for us given that previous research has shown that in some age groups up to 33.7% of participants are diagnosed with some type of Anxiety Disorder (Chou, 2010).

Some researchers have found a correlation be-

[1] https://www.dailystrength.org/

Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 37–42,
Austin, TX, November 5, 2016.

tween the usage of words related to anxiety with daily negative emotions (Tov et al., 2013). Applying our proposed framework, we found that the usage of words related to anxiety by active users in an online health support group has a steady decrease over the course of a user's involvement in the community and we theorize their daily negative emotion reduces as well. Our proposed framework can be easily extended to other related conditions such as depression or eating disorders. In general, we believe that sociolinguistic characteristics in online health support forums is an exciting topic that can shed additional light on human behavior and on the design of social media systems.

2 Related Work

Online Health Forums and social media: Online health communities are a rich source of data for the research community as a whole. Some researchers have studied the potential and limitation of such data and how it can augment existing public health capabilities and enable new ones (Dredze, 2012). One of the major concerns is the credibility of the information. Other researchers have studied how to automatically establish the credibility of the user generated medical statements by analyzing linguistic clues (Mukherjee et al., 2014). Other researchers have focused on understanding abstinence from tobacco or alcohol use (Tamersoy et al., 2015) and on how to find early indications of Adverse Drug Reactions from online healthcare forums (Sampathkumar et al., 2014). In the online world, several of the largest online health community websites are: MedHelp (`www.medhelp.org`), PatientsLikeMe (`www.patientslikeme.org`), and Daily-Strength (`www.dailystrength.org`)

Sociolinguistic patterns in social media: Social media is very appealing to the study of sociolinguistic analysis of the users. In particular one of the main concerns is how much information is actually posted by the users to justify the study of such textual data. In (Park, 2012) the authors found evidence that people post about their treatment on social media. Some researchers have shown the predictive power of studying linguistic patterns of social media users in order to predict depression (De Choudhury and Gamon, 2013). Furthermore, other researchers have studied insights about diseases, such as analyzing symptoms and medication usage and have found a strong correlation with public health data (Passarella, 2011). Ofoghi et al. (2016) have created an emotion classification of microblog content in order to study the public mood and effectively utilize it as an early warning system for epidemic outbreak. Also, they analyzed the emotions in microblog content after outbreaks to validate their approach. Finally, Aman and Szpakowicz (2007) describe an emotion annotation task and study how the inter-annotator agreement. They show how difficult is the emotion annotation task, the inter-annotator agreement ranges between 0.6 to 0.79.

Anxiety disorders: Researchers estimate the percentage of adults with anxiety disorders to vary from 3.2% (Fuentes and Cox, 1997) up to 14.2% (Norton et al., 2012). Other researchers have suggested that up to 33% of the general population will develop "clinical significant anxiety disorder" at some time in their life (Barlow et al., 2002). Anxiety disorders are commonly associated with medical conditions such as thyroid disease, asthma and heart disease (Diala and Muntaner, 2003). Also, some conditions such as coronary heart disease, hypertension, and hypoglycemia can be worsened through anxiety (Hersen and Van Hasselt, 1992).

3 Dataset Description

We collected data from Daily-Strength, one of the largest online support groups with more than 500 active groups based on the physical and mental conditions of its users. Daily-Strength allows users to create profiles, maintain friends, and join various condition-related support groups. It serves as a resource for patients to connect with others who have similar conditions. Users in these support groups can either *create* a new thread on a new topic[2], or *reply*[3] to a thread that someone else has created.

In the current study, we selected the support groups that had the most vibrant communities based on the number of unique users, and the number of unique threads. We focused on support groups with more than 1,000 different users and more than 200

[2]The topics are curated by the system administrators.

[3]The website does not distinguish between a reply to a main thread and a reply to a reply.

Table 1: Dataset statistics

Characteristic	Value
Number of support groups	93
Number of unique users	193,354
Number of unique posts	10,612,830

original threads. At the time of our data crawl, 93 groups fulfilled this selection constraint. Some of the most active support groups in this list include: Acne, ADHD (Attention Deficit Hyperactivity Disorder), Alcoholism, Asthma, Back Pain, Bipolar Disorder, Bone Cancer, COPD (Chronic Obstructive Pulmonary Disease), and Fibromyalgia. We crawled all of the original posts, thread initiations, and all the user replies for these support groups from the earliest available post until March 25, 2015. The posts and replies were downloaded as HTML files, one per thread, where each thread contains an initial post and zero or more replies. We filtered out posts from administrators of the website since they do not reflect the user's activities but just general guidelines or advice for the users.

Researchers have shown that a large and increasing number of people are going online for medical information and advice (Fox and Duggan, 2013). We want to base our analysis on a group of users who are consistently involved in the forum. Moreover, since we're interested in exploring if participation in the forum has any effects on its users and if these effects are reflected in linguistic patterns, it's important to analyze data from user posts across a significant period of time. We believe one year of activity will fulfill this purpose. Thus in subsequent analyses we filtered by users whose first and last post are at least one year apart, and who posted at least 50 posts during that interval. This filtering reduced the total number of users to approximately 10,000 users, still a significant number for our studies.

4 Analysis framework

We defined user behavioral dimensions (BDs) based on the word list provided by the Linguistic Inquiry and Word Count (LIWC) lexicon (Tausczik and Pennebaker, 2010). Each BD is defined as the corresponding word list from the LIWC lexicon. The LIWC contains 4,500 words and word stems, and each word or word stem defines one or more word categories or subdictionaries. This is a well developed tool aimed at revealing our thoughts, feelings, personality, and motivations based on our word usage. We focused on the anxious word list, but our analysis can be easily extended to any of the other LIWC list. Based on the emission of words in the linguistic dimensions in LIWC we created what we call user behavioral dimensions (BD). In order to study how the users of anxious words relate to other BDs defined by the LIWC we present in subsection 4.3 the correlation with six other BDs.

4.1 User's sub-population selection

The first part of our framework consists of creating a methodology to study a specific user's behavioral dimension. This methodology will help us find subpopulations corresponding to a particular BD. For this we propose metric for each of the BDs. We quantify each user BD according to:

$$BD_i(u) = \log\left(\frac{1}{|posts(u)|}\sum_{p\in posts(u)}\frac{|words_{BD_i}(p)|}{|words(p)|}\right)$$

where $i \in \{1,\ldots,N\}$ indexes the BDs, N is the number of different BDs, $posts(u)$ is the list of posts for user u, $words(p)$ is the list of words in post p, and $words_{BD_i}(p)$ is the list of words in post p that were on the list BD_i. In essence, we measure the average fraction of words from the list BD_i across the posts of a user. Since these fractions are less than 1 and we are using a logarithmic scale, $BD_i(u)$ will always have a negative magnitude.

We propose to base our analysis on the study of the extreme populations. To do that, we create a histogram which allows us to cluster users at the tails of the distribution. We create 4 clusters per condition out of this histogram. The lower 15% and 30% users; the top 30% and 15%. Figure 1 shows the generated histogram for the anxious word list from the LIWC. In order to select each population we sort the users according to the quantified values of the BD and then select the upper/lower part of the users.

4.2 Anxious word usage

We present the methodology used to analyze the BD corresponding to anxious words. The other BDs can be analyzed in an analogous way. The methodology we developed to analyze anxious word usage

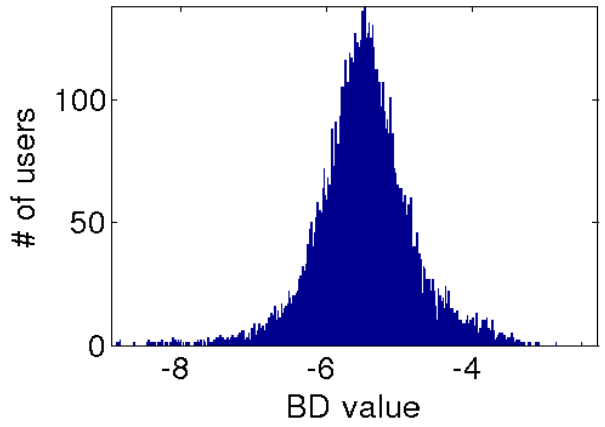


Figure 1: Sample usage of anxious related words.

is as follows. We first group all the posts based on the month when the user posted. All posts from the users during their first month, second month, third month, and so on. We then analyzed the usage pattern of anxious words over time. Figure 2 shows that the anxiousness of the users decreases in a constant manner over the course of their active involvement in the forum. This is in agreement with the effect of off-line support groups (Cain et al., 1986).

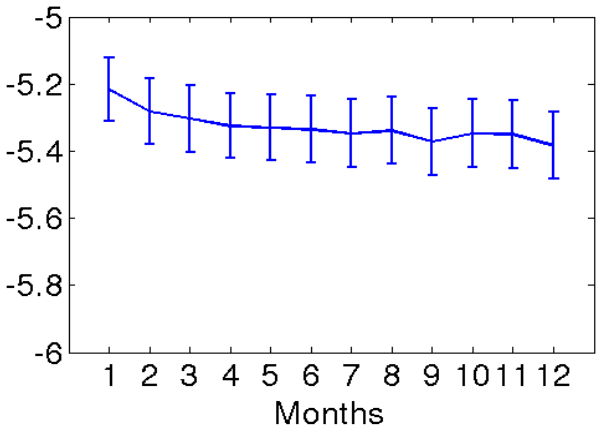


Figure 2: Single BD, anxious word usage.

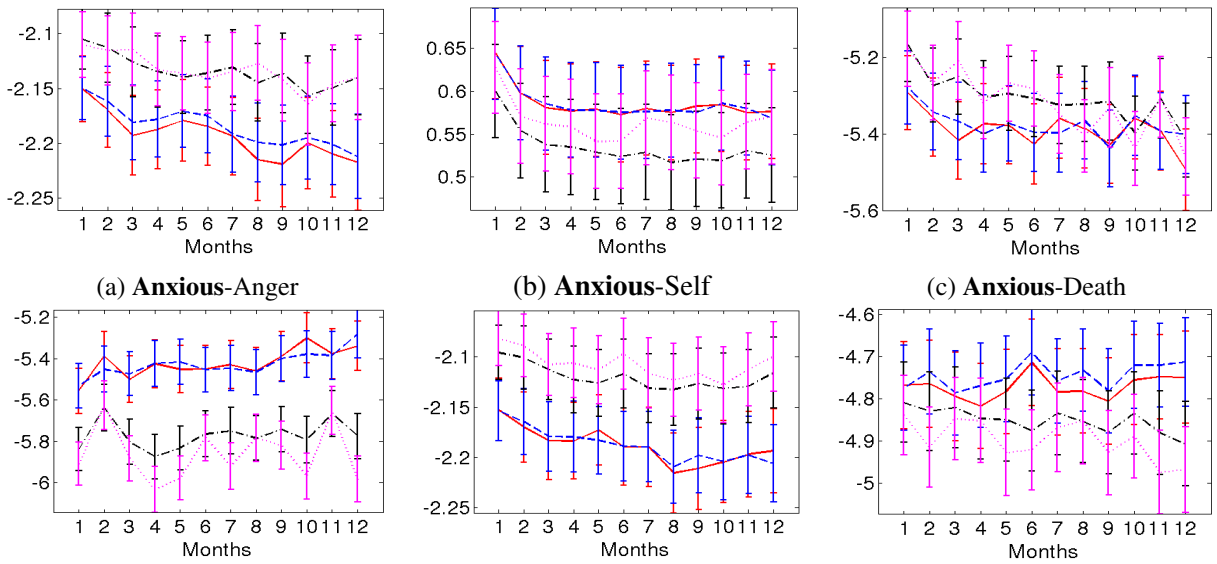
4.3 Correlation of anxious words with other word usage

In this part we focus on the analysis of BDs that correlate the most with anxious word usage. We ranked the BDs based on the correlation with anxious words usage. We used the Pearson coefficient correlation. In descending order the most correlated BDs are: a) anger, b) self-pronouns, c) death, d) money, e) present, and f) body. We present the results of all the correlations for completeness and describe the ones we found most interesting. For each of the clusters of users we identified on the anxious word usage, we plot the secondary BD score over time in Figure 3.

Figure 3 panels (a), (b), (c), and (e), shows that the usage of anger, self, death, and present words also decreases over time for all the low and high heavy users of anxious words. However from panels (d) and (f) we see that there is an increase of that particular word usage for some of the groups and a decrease from other groups.

Figure 3 (c) illustrates this for the case of anchor usage of anxious words and the figure shows the usage pattern of death words by those groups of users. We can see that users who used more anxious words (magenta dotted and black dash-dot) consistently used more death words over time than users who used fewer anxious words (red line and blue dashed). We theorize that this pattern can be related to suicidal topics similar to what other researchers have reported for depressed and anxious patients (Pompili et al., 2012)

From Figure 3 panel (d) we can see a difference between the amount of money related words used by people who use anxious words. The groups of users who use less anxious words (red line and blue dashed) tend to use more money related words, whereas the groups of users who use more anxious related words (magenta dotted and black dash-dot) tend to use less money words. In general researchers have linked people being tight with money as having more negative emotions such as been more anxious (McClure, 1984), however to the best of our knowledge this is the first study of such relation into a health support context where money is probably not directly related to how wealthy or not an individual is. From Figure 3 panel (e) we can see that people who use more anxious words (magenta dotted and black dash-dot) tend to use more present words, whereas people who use less anxious words (red line and blue dashed) use less present words. Some researchers have previously linked Defensive Pessimism people with being anxious in the present (Norem and Smith, 2006), we theorize that it can be a general pattern of people participating in online support forums.



(a) **Anxious**-Anger (b) **Anxious**-Self (c) **Anxious**-Death

(d) **Anxious**-Money (e) **Anxious**-Present (f) **Anxious**-Body

Figure 3: Correlation of the different BDs. Anchor BD anxious word usage (in boldface). Red line (-), blue line (–), black (-.), and magenta (..) corresponds to the lower 15%, 30% and upper 30%, 15% of the anxious word usage distribution.

5 Conclusion and Future Work

Similar to what other researchers (Cain et al., 1986) have shown for off-line support groups, based on our proposed framework for analyzing linguistic patterns of users of online support groups we conclude that the anxiety levels of patients involved in support groups lowers over time. We also conclude that anxiety levels are not directly related to money related talks in online support forums participants.

In this paper we have presented the correlation between Anxiety and BDs that we think are more interesting to study and the ones which are more relevant given the literature on Anxiety. However, a more detailed and robust method is been developed in order to rank the most relevant BDs which exhibit significant correlation over time.

References

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer.

David H Barlow, Susan D Raffa, and Elizabeth M Cohen. 2002. *Psychosocial treatments for panic disorders, phobias, and generalized anxiety disorder*, volume 2.

Eileen N. Cain, Ernest I. Kohorn, Donald M. Quinlan, Kate Latimer, and Peter E. Schwartz. 1986. Psychosocial benefits of a cancer support group. *Cancer*, 57(1):183–189.

Kee-Lee Chou. 2010. Panic disorder in older adults: evidence from the national epidemiologic survey on alcohol and related conditions. *International journal of geriatric psychiatry*, 25(1-3):822–832.

Munmun De Choudhury and Michael Gamon. 2013. Predicting depression via social media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2:128–137.

Chamberlain C. Diala and Carles Muntaner. 2003. Mood and anxiety disorders among rural, urban, and metropolitan residents in the United States. *Community Mental Health Journal*, 39(3):239–252.

Mark Dredze. 2012. How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84.

Susannah Fox and Maeve Duggan. 2013. Pew Internet & American Life Project. online http://www.pewinternet.org/2013/01/15/health-online-2013/.

Karina Fuentes and Brian J Cox. 1997. Prevalence of anxiety disorders in elderly adults: A critical analysis. *Journal of Behavior Therapy and Experimental Psychiatry*, 7916(4):269–279.

Michel Hersen and Vincent B. Van Hasselt. 1992. Behavioral assessment and treatment of anxiety in the elderly. *Clinical Psychology Review*, 12(1984):619–640.

Robert F. McClure. 1984. The relationship between money attitudes and overall pathology. *Psychology: A Journal of Human Behavior*, 21(1):4–6.

Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. 2014. People on drugs: credibility of user statements in health communities. *KDD '14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*.

JK. Norem and S. Smith. 2006. Defensive pessimism: positive past, anxious present, and pessimistic future. In *Judgements over time: The interplay of thoughts, feelings, and behaviors*, pages 34–46.

Joanna Norton, Maria Laure Ancelin, Rob Stewart, Claudine Berr, Karen Ritchie, and Isabelle Carriere. 2012. Anxiety symptoms and disorder predict activity limitations in the elderly. *Journal of Affective Disorders*, 141:276–285.

Bahadorreza Ofoghi, Meghan Mann, and Karin Verspoor. 2016. Towards early discovery of salient health threats: A social media emotion classification technique. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 21, page 504.

Minsu Park. 2012. Depressive moods of users portrayed in twitter. *ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*, page 18.

Ralph J Passarella. 2011. You are what you tweet : Analyzing twitter for public health. *In Empirical Natural Language Processing Conference (EMNLP).*, 56(8):1–12.

Maurizio Pompili, Marco Innamorati, Zoltan Rihmer, Xenia Gonda, Gianluca Serafini, Hagop Akiskal, Mario Amore, Cinzia Niolu, Leo Sher, Roberto Tatarelli, Giulio Perugi, and Paolo Girardi. 2012. Cyclothymic-depressive-anxious temperament pattern is related to suicide risk in 346 patients with major mood disorders. *Journal of affective disorders*, 136(3):405–11.

Hariprasad Sampathkumar, Xue-wen Chen, and Bo Luo. 2014. Mining adverse drug reactions from online healthcare forums using hidden Markov model. *BMC medical informatics and decision making*, 14(1):91.

David A. Sbarra. 2006. Predicting the onset of emotional recovery following nonmarital relationship dissolution: Survival analyses of sadness and anger. *Personality and Social Psychology Bulletin 32*, 3:298–312.

Acar Tamersoy, Munmun De Choudhury, and Duen Horng Chau. 2015. Characterizing smoking and drinking abstinence from social media. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 139–148. ACM.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

William Tov, Kok Leong Ng, Han Lin, and Lin Qiu. 2013. Detecting well-being via computerized content analysis of brief diary entries. *Psychological assessment*, 25(4):1069–78.