# Exploring Self-Identified Counseling Expertise in Online Support Forums

**Allison Lahnala**[1], **Yuntian Zhao**[1], **Charles Welch**[1], **Jonathan K. Kummerfeld**[1], **Lawrence An**[2,4], **Kenneth Resnicow**[3,4], **Rada Mihalcea**[1], **Verónica Pérez-Rosas**[1]
[1]Computer Science & Engineering, University of Michigan
[2]Medical School, University of Michigan
[3]School of Public Health, University of Michigan
[4]Center for Health Communications Research, University of Michigan
{alcllahn,clzhao,cfwelch,jkummerf,lcan,
kresnic,mihalcea,vrncapr}@umich.edu

## Abstract

A growing number of people engage in online health forums, making it important to understand the quality of the advice they receive. In this paper, we explore the role of expertise in responses provided to help-seeking posts regarding mental health. We study the differences between (1) interactions with peers; and (2) interactions with self-identified mental health professionals. First, we show that a classifier can distinguish between these two groups, indicating that their language use does in fact differ. To understand this difference, we perform several analyses addressing engagement aspects, including whether their comments engage the support-seeker further as well as linguistic aspects, such as dominant language and linguistic style matching. Our work contributes toward the developing efforts of understanding how health experts engage with health information- and support-seekers in social networks. More broadly, it is a step toward a deeper understanding of the styles of interactions that cultivate supportive engagement in online communities.

## 1 Introduction

Online social media forums play a critical role in health-related information sharing (Record et al., 2018). Health experts have noted that they can help reduce healthcare inequalities and improve access to health care, for instance by empowering coalitions of people living with chronic illness or specific disabilities (Griffiths et al., 2012), or by providing an anonymous forum for people seeking emotional support (De Choudhury and De, 2014). On the other hand, these forums elevate concerns about spreading medically inaccurate, misleading, or unsound information (Domínguez and Sapiña, 2015; Gage-Bouchard et al., 2018), which has had harmful public health impacts (Poland et al., 2011; Nobles et al., 2019). One study concluded that health information seekers in forums such as Reddit are likely to enact suggested behaviors regardless of perceived credibility (Record et al., 2018). However, the researchers also noted that this openness to information could be an opportunity for experts to encourage healthy behaviors through information sharing.

In this landscape, it is critical to understand the dynamics that cultivate safe communities that benefit the health and well-being of their participants and the broader implications for health communication (Chou et al., 2009). Health experts are thus considering social media's role in their interactions with patients and broader public health issues, and their role in engaging with the platforms (Domínguez and Sapiña, 2015; Nobles et al., 2019, 2020). This motivates an important research direction: understanding how experts engage with users in online platforms. This can inform platform design, moderation decisions, and health promotion efforts by experts.

This work focuses on understanding the engagement with professionals in the domain of mental health with two main research questions: (RQ1) Do experts have distinct influences as compared to non-experts in their interactions with support-seekers in online mental health?; and (RQ2) Do the experts' behaviors reflect established counseling principles and findings regarding behaviors associated with positive counseling outcomes? To answer these questions, we analyze responses from self-identified mental health professionals (MHP) to support-seekers in mental health and support communities on Reddit, and compare them to responses from other users who we refer to as *peers*. This is an important comparison, as many peers share similar health experiences, which prior work has found is associated with higher empathic concern (Hodges et al., 2010).

First, we test whether a text classifier can distin-



*Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4467–4480
August 1–6, 2021. 

guish between responses to support seekers from MHPs and peers. We find that it can, with 70% accuracy (well above random chance of 50%). Second, we analyze comments leading to further engagement with the support-seeking posters, as existing counseling principles emphasize the importance of eliciting client engagement in expert counseling sessions (Miller and Rollnick, 2012; Pérez-Rosas et al., 2018). Third, we analyze the users' linguistic tendencies, drawing inspiration from analyses of counseling conversations, which have offered insight into counselor behaviors associated with high quality sessions grounded in existing theories from psychology and counseling research using computational methods (Althoff et al., 2016; Pérez-Rosas et al., 2018; Zhang et al., 2019; Miller and Rollnick, 2012).

The main contributions of this work are: (1) We construct a dataset of mental health conversations from Reddit users with self-identified counseling expertise, covering a set of mental health subreddits annotated with categories denoting the type of mental health concern; (2) We develop a classifier that can distinguish between the language of MHPs and that of peers; (3) We perform an analysis of the differences in language use between MHPs and peers; and (4) We provide insight into language that leads to further engagement with support-seekers, comparing responses to peers and MHPs.

## 2 Related Work

Studies within the education and health domains have shown that advice and help-seeking interactions in online communities contribute positively to users' well-being, learning, and skills development (Campbell et al., 2016; Wang et al., 2015). This is particularly true for applications such as computer programming, career development, mentoring, coping with chronic or life-threatening diseases, and mental health issues (Baltadzhieva and Chrupała, 2015; Tomprou et al., 2019; Wang et al., 2015; De Choudhury and De, 2014).

In the mental health domain, studies have explored online support communities and many have found positive outcomes associated with anonymity, perceived empathy, and active user engagement (De Choudhury and De, 2014; Rheingold, 1993; Hodges et al., 2010; Welbourne et al., 2009; Nambisan, 2011). Computational approaches have aided studies in mental health forums, helping reveal positive relationships between linguistic accommodation and social support across subreddits (Sharma and De Choudhury, 2018). One example of insights from this work is that topic-focused communities like subreddits may enable more peer-engagement than non-community based platforms (Sharma et al., 2020). Other studies have revealed certain trade-offs of online support platforms, such as disparities in the level of support offered toward support-seekers of various demographics (Wang and Jurgens, 2018; Nobles et al., 2020) and in condolences extended across different topics of distress (Zhou and Jurgens, 2020). Studying MHP behaviors in such scenarios might help develop approaches that balance these trade-offs.

Computational approaches applied in these forums have also shed light on population-level health trends and health information needs, with examinations into how depression and post-traumatic stress disorder (PTSD) affect different demographic strata (Amir et al., 2019). Data mining has also been applied to understand adverse drug reactions (Wang et al., 2014) and public reactions towards infectious diseases (Park and Conway, 2017). Nobles et al. (2018) highlighted the potential for these forums to aid targeted health communication, for example by sharing information in r/STD, a subreddit about sexually transmitted diseases. Another case study of r/STD revealed the prevalence of diagnoses requests, and suggested that health professionals could partner with social media platforms to positively influence crowd-sourced diagnoses and help mitigate harmful misdiagnoses (Nobles et al., 2019). Record et al. (2018) found that health information seeking Reddit users are likely to enact suggested behaviors regardless of perceived credibility, providing further reason for health expert engagement to intervene when harmful information sharing occurs and promote healthy behavior.

Fewer studies have analyzed expert interactions in online forums. A study in a large Q&A community found that experts are more likely to provide help than peers and that their participation in discussions resulted in increased length and substance of discussions (Procaci et al., 2017). Recent studies have compared interactions with experts to interactions with peers in broader scientific communities (Park et al., 2020) and r/AskDocs on Reddit (Nobles et al., 2020). The latter paper closely relates to our study, as they also consider posts from experts on Reddit, but solely within r/AskDocs about different health topics and with

| | |
|---|---|
| Subreddits | 77 |
| Posts | 12,140 |
| Poster Replies | 24,357 |
| MHPs | 283 |
| Peers | 56,701 |
| Comments | |
| MHP | 9,685 |
| Peer | 92,698 |
| Total | 102,383 |
| Thread Length | |
| Mean | 8.4 |
| Median | 4 |
| Max | 64 |

Table 1: Dataset statistics.

| | |
|---|---|
| **Post:** | u/peer_user_X |
| I've recently been struggling with paranoid thoughts, for which I was hospitalized for my own safety. I do not feel suicidal anymore, however everyday is a long struggle of thinking everyone is an undercover agent out to get me or keep tabs on what I'm doing. I was hoping to hear some tips and stories if anyone else has dealt with similar thoughts and overcome them? Or are they something I will have to deal with for the rest of my life? Thanks in advance | |
| **Comment:** | u/MHP_user (LPC) |
| Paranoid thoughts are scared thoughts, justified or not. If you ignore the specific content of the thoughts and focus on the emotional valence (scared), is there something you can do in those moments to feel safer? | |
| **Poster Reply:** | u/peer_user_X |
| That's a good way of thinking about the situations as they arise. I will try to do that | |

Table 2: Example of an initial post, a reply from an MHP with the flair *LPC* (Licensed Professional Counselor), and a reply from the original user.

users of varying demographics.

The insights discussed above motivate investigations into how health experts and other users promote scientifically sound advice and offer supportive responses to health information seekers in online forums. In this work, we aim to contribute additional insights into expertise influence in online mental health communities by studying the dynamics of the communication process between support seekers and support providers.

## 3 Data Collection

We seek to understand the tendencies of users with professional experience, and more specifically counseling expertise, when interacting with support-seekers in online mental health and support-related forums. In uncovering which tendencies are associated with expertise, we enable further investigation into their role in the social dynamics of online support-seeking interactions, and potential applications of insight-driven recommendations for moderators and users of these forums.

**Source.** We use Reddit for its quantity of publicly available interactions in communities called *subreddits* that discuss mental health issues. In addition, Reddit has a system that allows users to indicate their professional expertise (Reddit Flairs), which we use to identify a set of users with mental health professional background, identified as *MHPs* during our study. We obtained flairs from the r/psychotherapy subreddit,[1] a decision motivated by their reliability, as the moderators of this community allow comments and posts only by licensed therapy providers who may be asked to submit proof if concerns of falsely posing as a therapist arise.[2] Sample flair tags in this set are: Psychiatrist (sometimes accompanied by MD or DO), LPC (or Licensed Professional Counselor), LMFT (or Licensed Marriage and Family Therapist), PsyD (or Doctorate of Psychology).

We use an existing list of mental health subreddits from r/ListOfSubreddits[3] with additions from manual observations; all of the subreddits in our dataset with their number of comments are in Appendix B in Table 5. From these, we retrieve threads where an MHP submitted a direct reply. During this step, we also kept posts made by peers i.e., individuals who did not use any of the mental health care professional flairs. Our collection spans threads created between November 29, 2009 and December 21, 2020. Table 1 shows descriptive statistics for the final composition of the dataset, and Table 2 shows a sample interaction demonstrating the structure we use for our analysis. This study focuses on direct replies to the poster, thus we attempt to eliminate *megathreads* which tend not to focus in individual support-seekers by removing those above the 95th percentile in their number of direct replies; we leave analysis of deeper nested replies for future work.

**Health Topics.** To understand whether particular topics influence interactions with support-seekers, we group the subreddits into broader topics based

[1] Degree and license flair descriptions from r/psychotherapy wiki.

[2] See rule 2 and 9 in https://www.reddit.com/r/psychotherapy/, also listed in Appendix A.

[3] r/ListOfSubreddit's compilation of mental health and advice subreddits.

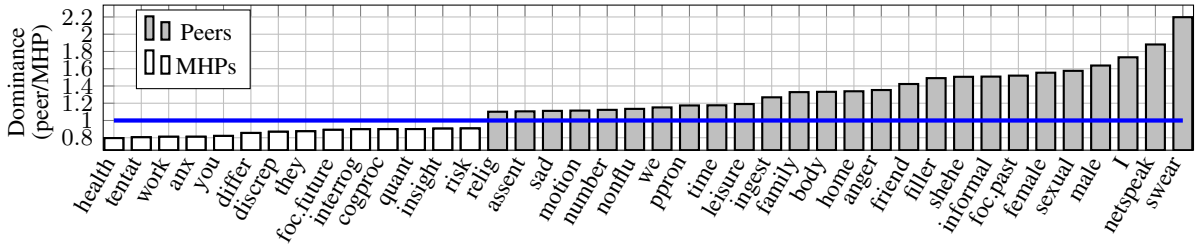


Figure 1: LIWC category dominance scores, computed as the relative use by peers divided by the relative use by MHPs, so that equal use is at $y = 1$ (blue line), higher dominance by peers at $y > 1$ (grey bars) and higher dominance by MHPs at $y < 1$ (white bars). Showing categories where frequency of use differs by at least 10%.

| Key | Topic |
|---|---|
| Trauma | Trauma & Abuse |
| Anx | Psychosis & Anxiety |
| Compuls. | Compulsive Disorders |
| Cope | Coping & Therapy |
| Mood | Mood Disorders |
| Addict. | Addiction & Impulse Control |
| Body | Eating & Body |
| Neurodiv. | Neurodevelopmental Disorders |
| Health | General |
| Social | Broad Social |

Table 3: Health condition and other subreddit topics. Keys are shortened names we use to refer to the topics.

on related health domains. We begin by following the categorization of subreddits by Sharma and De Choudhury (2018), who used the $k$-means clustering algorithm to generate initial clusters on the $n$-grams ($n = 3$) of the posts and manually refined the categories based the community descriptions in their subreddit home pages. Next, we adjust the categories and their associated subreddits based on the World Health Organization's ICD-10 classification system of mental and behavioural disorders.[4] The resulting topic categories are listed in Table 3 alongside shortened names which we use to refer to them. The full list of subreddits assigned to each topic are listed in Appendix B in Table 6.

## 4 Distinguishing MHPs and Peers

To begin our investigation into the linguistic behaviors of MHPs and peers, we test whether simple text classifiers are able to distinguish between comments authored by either MHPs or peers. We build three classifiers with different feature sets; the first are unigram counts for unigrams occurring at least five times, the second includes counts for the 73 word classes in the LIWC (Linguistic Inquire and Word Count) lexicon (Pennebaker et al., 2015), and the third encodes a subset of LIWC word classes associated with perspective shifts (i.e., *focusfuture, focuspast, focuspresent, I, ipron, negemo, posemo, ppron, pronoun, shehe, they, we,* and *you*) (Althoff et al., 2016); we elaborate on the psychological meaning behind these features in our analyses in the next section.

Due to the class imbalance between the peer and MHPs classes, we first downsampled the peer class to get a balanced distribution with the MHP class. This resulted in a set of 9,685 instances per class. We conduct our evaluations using ten-fold cross validation. Across these folds, the number of features ranges from 8,668 to 8,703. We use a Naive Bayes model, implemented with Sklearn's MultinomialNB module, [5] which outperformed a logistic regression model and an SVM in preliminary experiments.[6]

All models outperform a random baseline[7] with all LIWC features bringing the accuracy to 59.12%, LIWC perspective features to 59.14%, and unigram features to 70.80%. Overall, the classification results indicate language differences exist between the MHPs and peers. Motivated by this result, we proceed to several analyses to gain insights.

## 5 Linguistic and Dialogue Analysis

We analyze the linguistic behaviors of MHPs and peers responding to support-seeking posts, and their potential influence in eliciting further engagement with the support-seeker. Our analyses are inspired by psychology and computational studies that have shown that conversational behavioral aspects such as word usage, client engagement, and

[4] https://www.who.int/substance_abuse/terminology/icd_10/en/

[5] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

[6] Runs in ∼40 seconds per fold on one AMD Ryzen 7 3700U CPU.

[7] $p < 0.0001$ using a permutation test (Dror et al., 2018)

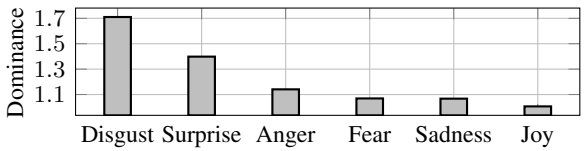


Figure 2: WordNet Affect usage (peers / MHPs)

language matching are positively related to successful counseling interactions (Gonzales et al., 2010; Althoff et al., 2016; Pérez-Rosas et al., 2018; Zhang et al., 2019).

### 5.1 Linguistic Ethnography

Numerous studies have demonstrated relationships between the dominant usage of certain word categories with individuals' psychological and physical health (Tausczik and Pennebaker, 2010; Weintraub, 1981; Rude et al., 2004). In alignment with these studies, we investigate the usage of such word categories using the LIWC and WordNet-Affect lexicons (Pennebaker et al., 2015; Strapparava and Valitutti, 2004).

For each group of users, we first compute the proportion of their words that fall in each category. Then, we compute the dominant use by dividing the proportion for peer users over the proportion for MHPs (Mihalcea and Pulman, 2009). Figure 1 shows LIWC categories where the rate of use differs by at least 10%, and results for WordNet Affect categories are shown in Figure 2.

Some observations such as the higher dominance of swear words (*swear*) and internet speak (*netspeak*) might be expected if professionals avoid such language. An interesting contrast in peers' language is the dominant use of first-person pronouns (*I, we*) and focus on the past (*focuspast*). In contrast, MHPs seem to use more non-first person pronouns (*you, they*) and focus on the future (*focusfuture*) instead. Peers' use of first person pronouns might arise when they share similar experiences with support-seekers. MHPs' use of second-person pronouns might suggest they are focusing on the support-seekers' experiences as a counselor would with a client in a counseling encounter. We also observe higher dominance of all WordNet Affect categories among peers, however the *joy* category (the most positive), is nearly equal with MHPs.

These observations of the peers' language are compelling because they align with existing theories linking depression to negative views of the future (i.e., *focuspast* and negative WordNet affects) (Pyszczynski et al., 1987) and self-focusing style (i.e., first-person pronouns) (Pyszczynski and Greenberg, 1987; Campbell and Pennebaker, 2003). Likewise, clients of SMS-based crisis counseling conversations were more likely to report feeling better after the encounter if they exhibited perspective shifts from these categories to their counterparts (i.e., toward *focusfuture*, non-first person pronouns, and positive sentiment) (Althoff et al., 2016).

Interestingly, the same study found clients were more likely to shift perspective when their counselors exhibited use of the counterpart categories first, suggesting that the counselors may play a key role in helping drive the perspective shift. Given those positive outcomes, observing the same dominant linguistic aspects among MHPs is encouraging and potentially signals a connection between how counselors apply conversational behaviors in practice and in online forum interactions. Future work can investigate the progression of dialogue between MHPs and support-seekers to find if support-seekers similarly exhibit the perspective shifts associated with the positive outcomes of the prior study, and likewise whether users of the forums also experience positive outcomes where this occurs.

### 5.2 Engaging Support-Seekers

To understand if linguistic behaviors are associated with prompting further engagement with the support-seeker, we compare the dominance of LIWC categories in comments receiving replies compared to comments that do not by dividing the usage rates of the former by the latter. Figure 3 shows these ratios for categories that differ by at least 5%. A compelling observation is the dominance of the categories *health*, *tentat*, and *you* in the MHP comments prompting poster-replies, and *you*, *focusfuture*, *interrog*, and *health* in the peer comments prompting poster-replies, as was exhibited among MHPs (see Figure 1 in Section 5.1); on the other end, the categories are more dominant in comments that do not engage a reply such as *I*, *we*, *death*, *friend*, *relig*, *swear*, were similarly represented as dominant categories in the peer group.

To gain further insight into these observations, we perform the following analysis: for each user group (peers and MHPs) we use a foreground corpus of their comments that were replied to by the support-seekers, and a background corpus of their comments that were not, and compute the dominance of LIWC categories of the foreground over

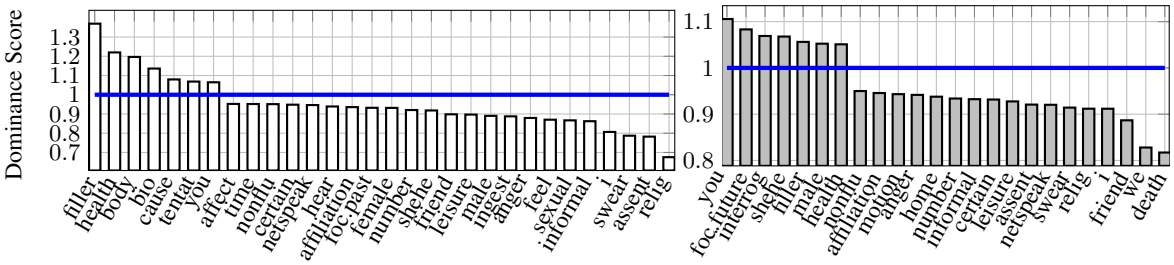


Figure 3: Dominance of LIWC categories, computed as the category relative frequencies among comments that **prompt support-seeker responses** divided by the relative frequencies among comments that do not, computed separately for MHPs (left) and peers (right).

| DRR Group | OR Group | $\tau$ | p-value |
|---|---|---|---|
| Peer | MHP | .191 | .017 |
| MHP | MHP | .158 | .048 |
| Peer | Peer | -.031 | .689 |
| MHP | Peer | .008 | .916 |

Table 4: Kendall $\tau$'s coefficient between the LIWC category dominance ranking in the replied comments (DRR) of the user group on the left and the overall ranking of LIWC category usage (OR) by the user group to the right.

the background as a ratio of their relative frequencies. We then rank the categories by highest to lowest dominance scores, and refer to this ranking by DRR (for **D**ominance **R**ank for **R**eplied comments). We compare the DRRs of each user group to the ranking of LIWC category usage among MHP users and among peer users separately (from Section 5.1) by computing the Kendall Tau's coefficient between them. A positive correlation would thus indicate that the more (or less) dominant categories among a group's replied comments are also more (or less) dominant among the other group overall. The correlation coefficients are shown in Table 4.

Interestingly, we observe a slight positive correlation between the DRRs for both MHPs and peers with the overall LIWC category usage ranking for MHPs. On the other hand, we see no correlations with the LIWC usage ranks for peers. Intuitively, it appears that for both MHPs and peers, the comments prompting further engagement with the poster appear to reflect the overall dominant linguistic aspects captured by LIWC of MHPs, but not peers. As counseling principles have emphasized the importance of mutual engagement between counselors and clients (Miller and Rollnick, 2012) and other work has shown that higher quality counseling sessions are associated with higher client engagement (Pérez-Rosas et al., 2018), it is compelling to observe associations between linguistic aspects of MHPs with the aspects associated with poster-engagement.

### 5.3 Linguistic Style Matching

Linguistic Style Matching (LSM) measures the extent to which one speaker matches another (Gonzales et al., 2010). It compares two parties' relative use of function words as these words are more indicative of style rather than content (Ireland and Pennebaker, 2010).

Previous studies in counseling conversations have measured LSM to understand the extent that counselors and clients match their language. Pérez-Rosas et al. (2019) showed higher LSM for high quality counseling sessions whereas Althoff et al. (2016) showed lower LSM for higher quality sessions. Pérez-Rosas et al. (2019) attributed this to the differences between the conversations they analyzed, theirs being synchronous face-to-face interactions while Althoff et al. (2016)'s was of asynchronous text messages, as well as differences in counseling styles.

We follow Nobles et al. (2020)'s approach leveraging Ireland and Pennebaker (2010)'s procedure to measure LSM between support seekers and support providers. For a text sequence, we compute the percentage of words that belong to each of nine function-word categories $c$ from the LIWC lexicon, which include *auxiliary verbs, articles, common adverbs, personal/impersonal pronouns, prepositions, negations, conjunctions,* and *quantifiers*. Then, we compute the LSM of each word category $c$ as shown in Equation 1 where $p$ represents *post* and $r$ represents the response. The composite LSM score for $p$ and $r$ is the mean of all category LSM scores. For each thread, we separate the MHP and peer replies, and take the mean of all

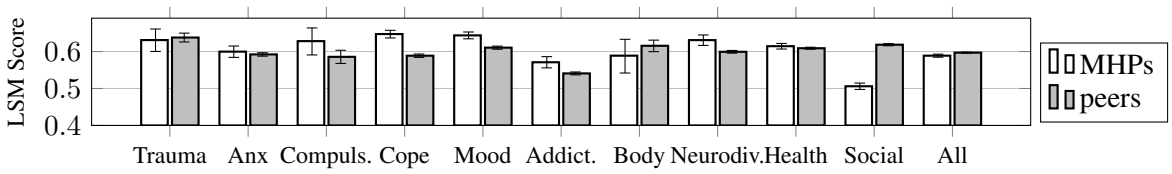


Figure 4: LSM scores with 95% confidence intervals calculated with non-parametric bootstrap resampling.

composite LSM scores.

$$LSM_c = 1 - \frac{abs(cat\%_p - cat\%_r)}{cat\%_p + cat\%_r + .0001} \quad (1)$$

We compute these LSM scores over all data together as well as separately for each subreddit topic (named in Table 3). The resulting scores are shown in Figure 4.

We observe LSM scores vary by topic, and most are similar for peers and MHPs or have overlapping confidence intervals. Compared to their LSMs in other topics, MHPs score lower in SOCIAL, which covers broad social issues that are less specialized to health conditions than the others. However, peers have high LSMs in SOCIAL relative to most other topics, and notably higher LSMs than the MHPs. Additionally, MHPs have higher LSMs than those of peers and relative to their own in communities that cover topics of specific compulsive, mood, and neurodevelopmental disorders (COMPULS., MOOD, and NEURODIV.), communities that orient toward counseling purposes (COPE), or toward advice-seeking communities for health and social concerns (HEALTH). The influences in these results require further investigation, but a possible explanation could be that expert knowledge and experience may offer more benefit to specialized condition-related issues than to broader social issues.

## 6 Language Modeling

We further examine differences in word usage by building separate language models for MHPs and peers. We seek to identify language use that is indicative of one group or another by running the language model of one on the data of the other and analyzing words with high perplexity. To run these experiments, we use the language model of Merity et al. (2018a,b), which is a recent LSTM-based language model that achieved state-of-the-art performance by combining several regularization techniques.[8]

Our implementation uses a fixed vocabulary of 20,907 tokens for both the peer and MHP language models. This is determined by a minimum count of five across the set of posts from both groups. Each language model is trained for 50 epochs.[9]

We use the language model trained on MHP data to find words with high entropy in peer data and vice versa. Since we are concerned with the *difference* in predictability of words between the MHP and peer language models, we subtract the entropy given by the model trained on that data from the entropy assigned by the model that was not trained on that data. In other words, to find words difficult to predict in B's data, we subtract each word's entropy calculated by the model trained on B from the entropy calculated by the model trained on A as follows, for a set of words, X:

$$E_{A,B} = -\frac{1}{|X|}\sum_{x \in X} log(p_A(x)) - log(p_B(x)) \quad (2)$$

If we calculate the entropy difference for each LIWC category and for each assignment of the MHP and peer groups to A and B, we find the highest differences for each category shown in the first and third plots of Figure 5. We find highest entropy scores for words relating to *leisure*, *sex*, and *numbers* when running the MHP language model on peer data. Likewise, when running the peer model on MHP data, the category of *discrepancy* contains words whose accuracy is improved the least by the peer model, again showing that these words are more indicative of the MHP group.

We perform a similar analysis, creating a language model for posts which have the highest score (or tied for highest) and another model for all other

[8] https://github.com/salesforce/awd-lstm-lm

[9] Validation set perplexities for expert and score groups: peer on peer: 44, peer on MHP: 52, MHP on peer: 91, MHP on MHP: 74, low on high: 39, low on low: 43, high on high: 50, high on low: 57. The difference in perplexity is due to the difference in volume of posts between groups. Runs in ∼2 min per epoch on a GeForce RTX 2080 Ti GPU.

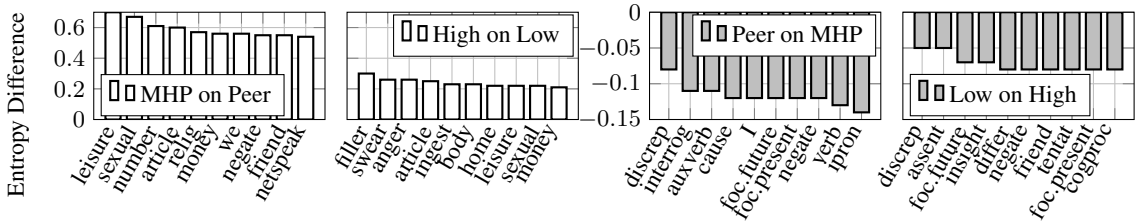


Figure 5: Entropy differences for LIWC word categories when running both language models on one group's data. High entropy scores on one dataset indicate word types that are harder for the opposite group's model to predict.

posts. We measure entropy differences and show the highest scoring categories for each group in the second and fourth plots. Some of the categories indicative of MHP language are also indicative of higher scoring posts; *discrepancy*, *present* and *future* words, and *negation* words, while other categories like *assent* and *insight* words are more dominant in higher scoring posts. The lower scoring posts have the highest entropy differences for some types of words in the peer data, however, we also see that *filler*, *anger*, and *swear* words had the highest entropy differences for the low scoring group. Qualitative example sentences with word-level entropy and LIWC annotations are shown in the appendix in Table 7.

## 7 Discussion and Future Work

In comparing linguistic aspects of MHP and peer comments, we find MHP tendencies align with established counseling principles and findings in counselor behaviors from recent literature. In particular, they align in the use of words that increase the likelihood of desired perspective shifts associated with clients feeling better after text counseling sessions (RQ2) (Althoff et al., 2016). We also found unique differences in the behavior of MHPs as compared to peers in how they respond to information seekers (RQ1). Although, comments by peers that prompt support-seeker replies also make use of similar word categories to MHPs, which shows that comparing MHPs to peers can offer insight into peer interactions as well.

It is important to note that our analyses rely heavily on the LIWC lexicon. While LIWC and other lexicons can help uncover variational language across groups at an exploratory stage, their use alone does not explain why variations are present. Certain limitations of LIWC are clear, such as when certain words that occur in multiple categories misleadingly boost the prominence of the categories equally. Kross et al. (2019)'s and Jaidka et al. (2020)'s studies have also demonstrated limitations of the use of LIWC when working with word counts to correlate with well-being metrics and an individual's emotional state. We utilize LIWC to understand linguistic behavior differences in conversations with peers and MHPs rather than to evaluate the emotional or mental health state of individuals; however, it is important to consider how these limitations could pertain to our interpretations of their differences, especially as we explore them more deeply in future work. In our study, we explore the patterns we find in the context of previous findings from related literature such as (Althoff et al., 2016) and (Nobles et al., 2020), however it warrants another study into nuanced aspects of the patterns to infer their social functions in support seeking forums in particular.

Although our findings align MHP behaviors with certain counselor behaviors associated with positive outcomes, our analyses do not support claims that MHP behaviors are more beneficial to individuals seeking support; rather, we have shown that the general tendencies of MHPs are in accordance with principles and behaviors demonstrated by counselors in other settings. Understanding the outcomes of these interactions for individual support seekers remains as an area for future work, which could employ surveying methods from prior work to measure perceived empathy in online communities (Nambisan, 2011). Our dataset also enables investigations into whether support-seekers exhibit perspective shifts in interacting with MHPs or peers, and what MHP and peer tendencies are associated with these perspective shifts.

Another direction for future work could focus on modeling social media-specific engagement patterns of MHP and peer interactions. Prior work developed a model that accounts for variables indicating the level of attention threads receive (i.e., thread lengths and number of unique commenters), and variables indicating the degree of interaction

between posters and commenters (i.e., time between responses and whether the poster replies to commenters), and used this model to study peer-to-peer interactions in online mental health platforms (Sharma et al., 2020); this approach may enable studying supportive interactions in megathreads and threads involving back-and-forth dialog between two or more parties.

More questions arise if we consider MHP tenure and specific domain of expertise (e.g., specializations, licenses, academic degrees). Prior work that studied longitudinal changes in counselor linguistic behaviors indicated that systematic changes occur over time as counselors develop personal styles that are more distinct from other counselors and exhibit more diversity across interactions (Zhang et al., 2019). Future work could model the language longitudinally for MHPs and peers that have longer-term histories of participating in mental health forums to investigate whether systematic changes occur online as well, and if so, whether they reflect similar changes found in prior work.

## 8 Limitations and Ethical Considerations

A number of unknowns exist in what we are able to extract from Reddit. For instance, we do not know if users that do not use flairs are mental health professionals. We assume that those who have used the MHP flairs are MHPs and those that have not used them are peers. Additionally, we have grouped all MHP flairs into one group for our analysis, though a more nuanced analysis based on particular professional roles (e.g., psychologists, psychiatrists, social workers) and specializations (e.g., motivational interviewing, cognitive behavioral therapy, family & marriage counseling) may reveal additional trends. Prior work found that disclosing credentials has impacts on engagements that vary by subreddit and linguistic patterns associated with different experience levels and expert domains (Park et al., 2020), thus the effects of disclosing MHP credentials when responding to support-seekers should be investigated.

A classifier or language model used to distinguish between MHPs and peers or to generate the language of either could have negative implications. A generative model that provides feedback to users could generate language that is harmful for those seeking help. Our work could be used to devise a tool to train counselors, however we do not have a direct measure of what type of responses are helpful or meaningful. In such an application, there is potential to reinforce harmful behaviors due to the inaccuracy of our models. Future studies are needed to determine how to best design a tool to train counselors and how models derived from corpora such as ours correspond to advice that patients find useful.

## 9 Conclusion

As the role of social networks is becoming more critical in how people seek health-information, it is important to understand their broader implications to health communication and how health experts can engage to promote the soundest information and offer support to their vulnerable users. By elucidating techniques employed by mental health professionals in their interactions with support-seekers in mental health forums, we have contributed insights toward the broader research direction of understanding how health experts currently engage with these platforms. With evidence that MHP linguistic behaviors associate with further engagement with support-seekers and that these same behaviors are associated with positive counseling conversation outcomes, we have shown that analyzing MHP behavior is a promising direction for better understanding online interaction outcomes, which can further inform forum design and moderation, and expert health promotion efforts.

The code used for our experiments and analyses, and the post ids in our dataset can be accessed at https://github.com/MichiganNLP/MHP-and-Peers-Reddit.

## Acknowledgments

This material is based in part upon work supported by the Precision Health initiative at the University of Michigan, by the National Science Foundation (grant #1815291), and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Precision Health initiative, the National Science Foundation, or John Templeton Foundation.

## References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to

mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Silvio Amir, Mark Dredze, and John W Ayers. 2019. Mental health surveillance over social media with digital cohorts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 114–120.

Antoaneta Baltadzhieva and Grzegorz Chrupała. 2015. Predicting the quality of questions on Stackoverflow. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 32–40, Hissar, Bulgaria. INCOMA Ltd. Shoumen, Bulgaria.

Julie Campbell, Cecilia Aragon, Katie Davis, Sarah Evans, Abigail Evans, and David Randall. 2016. Thousands of positive reviews: Distributed mentoring in online fan communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 691–704, New York, NY, USA. Association for Computing Machinery.

R Sherlock Campbell and James W Pennebaker. 2003. The secret life of pronouns: Flexibility in writing style and physical health. *Psychological science*, 14(1):60–65.

Wen-Ying Sylvia Chou, Yvonne M Hunt, Ellen B Beckjord, Richard P Moser, and Bradford W Hesse. 2009. Social media use in the united states: implications for health communication. *Journal of medical Internet research*, 11(4):e48.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Martí Domínguez and Lucía Sapiña. 2015. Pediatric cancer and the internet: exploring the gap in doctor-parents communication. *Journal of Cancer Education*, 30(1):145–151.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Elizabeth A Gage-Bouchard, Susan LaValley, Molli Warunek, Lynda Kwon Beaupin, and Michelle Mollica. 2018. Is cancer information exchanged on social media scientifically accurate? *Journal of cancer Education*, 33(6):1328–1332.

Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19.

Frances Griffiths, Jonathan Cave, Felicity Boardman, Justin Ren, Teresa Pawlikowska, Robin Ball, Aileen Clarke, and Alan Cohen. 2012. Social networks–the future for health care delivery. *Social science & medicine*, 75(12):2233–2241.

Sara D Hodges, Kristi J Kiel, Adam DI Kramer, Darya Veach, and B Renee Villanueva. 2010. Giving birth to empathy: The effects of similar experience on empathic accuracy, empathic concern, and perceived empathy. *Personality and Social Psychology Bulletin*, 36(3):398–409.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549.

Kokil Jaidka, Salvatore Giorgi, H Andrew Schwartz, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2020. Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, 117(19):10165–10171.

Ethan Kross, Philippe Verduyn, Margaret Boyer, Brittany Drake, Izzy Gainsburg, Brian Vickers, Oscar Ybarra, and John Jonides. 2019. Does counting emotion words on online social networks provide a window into people's subjective experience of emotion? a case study on facebook. *Emotion*, 19(1):97.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018a. An analysis of neural language modeling at multiple scales. *arXiv preprint arXiv:1803.08240*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018b. Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations*.

Rada Mihalcea and Stephen Pulman. 2009. Linguistic ethnography: Identifying dominant word classes in text. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 594–602. Springer.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Priya Nambisan. 2011. Information seeking and social support in online health communities: impact on patients' perceived empathy. *Journal of the American Medical Informatics Association*, 18(3):298–304.

Alicia Nobles, Caitlin Dreisbach, Jessica Keim-Malpass, and Laura Barnes. 2018. " is this an std? please help!": Online information seeking for sexually transmitted diseases on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Alicia L Nobles, Eric C Leas, Benjamin M Althouse, Mark Dredze, Christopher A Longhurst, Davey M Smith, and John W Ayers. 2019. Requests for diagnoses of sexually transmitted diseases on a social media platform. *Jama*, 322(17):1712–1713.

Alicia L Nobles, Eric C Leas, Mark Dredze, and John W Ayers. 2020. Examining peer-to-peer and patient-provider interactions on a social media community facilitating ask the doctor services. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 464–475.

Albert Park and Mike Conway. 2017. Tracking health related discussions on reddit for public health applications. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1362. American Medical Informatics Association.

Kunwoo Park, Haewoon Kwak, Hyunho Song, and Meeyoung Cha. 2020. "trust me, i have a ph. d.": A propensity score analysis on the halo effect of disclosing one's offline social status in online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 534–544.

James W Pennebaker, Roger J Booth, Ryan L Boyd, and Martha E Francis. 2015. Linguistic inquiry and word count: Liwc2015.

Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. Analyzing the quality of counseling conversations: the tell-tale signs of high-quality counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935.

Gregory A Poland, Robert M Jacobson, et al. 2011. The age-old struggle against the antivaccinationists. *N Engl J Med*, 364(2):97–9.

T. B. Procaci, S. W. M. Siqueira, B. P. Nunes, and T. Nurmikko-Fuller. 2017. Modelling experts behaviour in q a communities to predict worthy discussions. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, pages 291–295.

Tom Pyszczynski and Jeff Greenberg. 1987. Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychological bulletin*, 102(1):122.

Tom Pyszczynski, Kathleen Holt, and Jeff Greenberg. 1987. Depression, self-focused attention, and expectancies for positive and negative future life events for self and others. *Journal of personality and social psychology*, 52(5):994.

Rachael A Record, Will R Silberman, Joshua E Santiago, and Taewook Ham. 2018. I sought it, i reddit: Examining health information engagement behaviors among reddit users. *Journal of Health Communication*, 23(5):470–476.

Howard Rheingold. 1993. *The virtual community: Homesteading on the electronic frontier*, volume 32. Addison-Wesley Reading, MA.

Stephanie S Rude, Eva-Maria Gortner, and James W Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133.

Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020. Engagement patterns of peer-to-peer interactions on mental health platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 614–625.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA. Association for Computing Machinery.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1083–1086.

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.

Maria Tomprou, Laura Dabbish, Robert E. Kraut, and Fannie Liu. 2019. Career mentoring in online communities: Seeking and receiving advice from an online community. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Sheng Wang, Yanen Li, Duncan Ferguson, and Chengxiang Zhai. 2014. Sideeffectptm: An unsupervised topic model to mine adverse drug reactions from health forums. In *Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics*, pages 321–330.

Yi-Chia Wang, Robert E Kraut, and John M Levine. 2015. Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. *Journal of medical Internet research*, 17(4):e99.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45.

Walter Weintraub. 1981. *Verbal behavior: Adaptation and psychopathology*. Springer Publishing Company New York.

Jennifer L Welbourne, Anita L Blanchard, and Marla D Boughton. 2009. Supportive communication, sense of virtual community and health outcomes in online infertility groups. In *Proceedings of the fourth international conference on Communities and technologies*, pages 31–40.

Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding your voice: The linguistic development of mental health counselors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 936–947.

Naitian Zhou and David Jurgens. 2020. Condolences and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626.

# Appendix

## A Flairs

Rules regarding flair credibility from r/psychotherapy:

"**2. Only posts and comments from those providing therapy in a licensed capacity allowed.** No comments/posts from anyone who is not providing therapy in a licensed capacity. This includes students who are not yet practicing therapy (e.g., undergraduate or graduate students who haven't had their first practica experience) or if you have left the field for another field, this is not the place for you to post/comment. There is an exception to this rule for posting in our Career and Education Megathread. Accurate user flair is required for all posts, and strongly encouraged for comments."

"**9. Falsely posing as a therapist** If you post in this subreddit, the assumption is made that you are a therapist. Users that falsely post as if they were a therapist will be permanently banned. Claiming that you didn't say you were a therapist is not an argument against this rule. Users may be asked to submit proof of their status as a practicing therapist to appeal a ban."

## B Data

We used the PushShift API for the first pass of obtaining mental health posts and comments, and the MHP flairs. After extracting the IDs of posts where MHPs commented, we obtained the fully structured comment sections using open sourced code from https://github.com/saucecode/reddit-thread-ripper. The numbers of posts in our dataset for each subreddit are shown in Table 5.

## C Other

Sample sentences illustrating relative entropies of words predicted by the peer language model on MHP data (top) and the MHP language model on peer data (bottom) are shown in Table 7.

| | | | | | |
|---|---|---|---|---|---|
| AskDocs | 21025 | relationship_advice | 16061 | stopdrinking | 10170 |
| ADHD | 9093 | offmychest | 5076 | mentalhealth | 4486 |
| socialskills | 4113 | BPD | 3570 | depression | 3235 |
| Anxiety | 2956 | aspergers | 2703 | Advice | 2493 |
| askatherapist | 2120 | PCOS | 1819 | alcoholicsanonymous | 1498 |
| leaves | 1452 | SuicideWatch | 1092 | REDDITORSINRECOVERY | 977 |
| needadvice | 892 | ptsd | 704 | NoFap | 465 |
| OCD | 411 | socialanxiety | 400 | BipolarReddit | 355 |
| GetMotivated | 354 | alcoholism | 350 | cripplingalcoholism | 335 |
| emetophobia | 297 | bulimia | 249 | mentalillness | 246 |
| nosurf | 224 | EOOD | 208 | depression_help | 193 |
| EatingDisorders | 170 | schizophrenia | 167 | MMFB | 159 |
| AlAnon | 139 | disability | 127 | fuckeatingdisorders | 119 |
| Antipsychiatry | 116 | MadOver30 | 114 | quittingkratom | 114 |
| addiction | 111 | GFD | 109 | CompulsiveSkinPicking | 108 |
| Needafriend | 106 | dbtselfhelp | 99 | rapecounseling | 93 |
| stopsmoking | 89 | selfhelp | 87 | ForeverAlone | 81 |
| getting_over_it | 72 | BodyAcceptance | 54 | Anger | 50 |
| traumatoolbox | 50 | selfharm | 47 | TwoXADHD | 40 |
| survivorsofabuse | 40 | dpdr | 38 | rape | 36 |
| Tourettes | 34 | HealthAnxiety | 26 | schizoaffective | 25 |
| Anxietyhelp | 25 | eating_disorders | 20 | domesticviolence | 17 |
| neurodiversity | 13 | helpmecope | 12 | StopSelfHarm | 12 |
| sad | 11 | AtheistTwelveSteppers | 10 | Trichsters | 6 |
| MenGetRapedToo | 5 | ARFID | 5 | whatsbotheringyou | 3 |
| DysmorphicDisorder | 1 | OCPD | 1 | | |

Table 5: The number of comments in each subreddit of our dataset.

| Category | Subreddits |
|---|---|
| Trauma & Abuse (Trauma) | r/Anger, r/survivorsofabuse, r/domesticviolence, r/ptsd, r/rapecounseling, r/selfharm, r/StopSelfHarm, r/traumatoolbox, r/rape, r/MenGetRapedToo |
| Psychosis & Anxiety (Anx) | r/Anxiety, r/socialanxiety, r/Anxietyhelp, r/HealthAnxiety, r/BPD, r/dpdr, r/schizophrenia, r/schizoaffective, r/emetophobia |
| Compulsive Disorders (Compuls.) | r/CompulsiveSkinPicking, r/OCD, r/Trichsters, r/DysmorphicDisorder, r/OCPD |
| Coping & Therapy (Cope) | r/getting_over_it, r/helpmecope, r/offmychest, r/MMFB, r/askatherapist, r/EOOD, r/dbtselfhelp, r/AlAnon, r/REDDITORSINRECOVERY, r/GetMotivated, r/Antipsychiatry, r/selfhelp |
| Mood Disorders (Mood) | r/depression, r/depression_help, r/ForeverAlone, r/GFD, r/mentalhealth, r/SuicideWatch, r/sad, r/BipolarReddit |
| Addiction & Impulse Control (Addict.) | r/stopdrinking, r/addiction, r/stopsmoking, r/leaves, r/alcoholism, r/cripplingalcoholism, r/quittingkratom, r/alcoholicsanonymous, r/NoFap |
| Eating & Body (Body) | r/eating_disorders, r/EatingDisorders, r/ARFID, r/fuckeatingdisorders, r/BodyAcceptance, r/bulimia |
| Neurodevelopmental Disorders (Neurodiv.) | r/ADHD, r/aspergers, r/TwoXADHD |
| General (Health) | r/AskDocs, r/needadvice, r/Advice, r/mentalillness, r/neurodiversity, r/whatsbotheringyou, r/MadOver30 |
| Broad Social (Social) | r/socialskills, r/relationship_advice, r/nosurf, r/Needafriend, r/AtheistTwelveSteppers, r/PCOS, r/disability, r/Tourettes |
| Overall | All |

Table 6: Subreddit categories.

| | |
|---|---|
| | MHP Data |
| DISCREP | the problem with psychiatric research is the relative subjectivity of it , much less glamorous outcomes , and the lack of public interest despite its burden on society . |
| INTERROG | who diagnosed you with spinal issues - which might show how anxiety affects your physical health , |
| AUXVERB | that being said , many other mental health concerns have overlapping symptoms with adhd including anxiety and depression . |
| | Peer Data |
| LEISURE | it 's like jogging with a back back full of bricks . |
| SEXUAL | it seems like everybody here is desired sexually so that must mean you 're doing something right . |
| NUMBER | when it came time for homework she set them up with three things to do ( one being homework ) and had them switch every fifteen minutes . |

Table 7: Sample sentences from MHP data with relative entropy marked by highlight color (i.e. darker blue means higher entropy relative to other words in the sentence). All words in the given LIWC category are marked with a rounded rectangle.