Moments of Change: Analyzing Peer-Based Cognitive Support in Online Mental Health Forums

Yada Pruksachatkun[*]
NYU Center for Data Science
New York City, New York, USA
yp913@nyu.edu

Sachin R. Pendse[*]
Microsoft Research India
Bangalore, India
t-sapen@microsoft.com

Amit Sharma
Microsoft Research India
Bangalore, India
amshar@microsoft.com

ABSTRACT

Clinical psychology literature indicates that reframing irrational thoughts can help bring positive cognitive change to those suffering from mental distress. Through data from an online mental health forum, we study how these cognitive processes play out in peer-to-peer conversations. Acknowledging the complexity of measuring cognitive change, we first provide an operational definition of a "moment of change" based on sentiment change in online conversations. Using this definition, we propose a predictive model that can identify whether a conversation thread or a post is associated with a moment of cognitive change. Consistent with psychological literature, we find that markers of language associated with sentiment and affect are the most predictive. Further, cultural differences play an important role: predictive models trained on one country generalize poorly to others. To understand *how* a moment of change happens, we build a model that explicitly tracks topic and associated sentiment in a forum thread.

CCS CONCEPTS

• **Human-centered computing** → **Social network analysis**; • **Computing methodologies** → **Natural language processing**; **Discourse, dialogue and pragmatics**.

KEYWORDS

mental health; computational clinical psychology; social computing; social media; online communities; topic modeling

[*] Equal contribution

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK
© 2019 Association for Computing Machinery.
ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.1145/3290605.3300294

ACM Reference Format:
Yada Pruksachatkun, Sachin R. Pendse, and Amit Sharma. 2019. Moments of Change: Analyzing Peer-Based Cognitive Support in Online Mental Health Forums. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300294

1 INTRODUCTION

One of the most empirically supported and popular forms of psychotherapy [34], Cognitive-Behavioral Therapy (CBT) is grounded in the well observed phenomenon [13, 30] that people experiencing psychiatric distress are more prone to negative beliefs about their future, self, and world [11]. Additionally, these negative beliefs can serve as the base for the cognitive biases [12] that sustain a depressive episode. To counter these negative beliefs, a foundational part of CBT is the process of working alongside someone experiencing distress to restructure these automatic thoughts and reinterpret emotional stimuli, a process called *cognitive restructuring* [11]. The cognitive changes brought about by this process have been shown to have a strong relation with symptom change [8, 27] and therapeutic improvement in the treatment of depression [42]. As a result, cognitive change has been theorized as a causal mechanism of change in the treatment of depression and associated distress [45].

Cognitive restructuring is considered a particularly powerful exercise because it can be effective outside the bounds of the clinician's office. For instance, it has been administered successfully through the internet [19] and through mobile phones [69]. In particular, cognitive restructuring is also helpful in controlled settings when administered via peers in online support forums [38, 48, 52].

Correspondingly, many such online forums exist that facilitate *virtual* therapeutic interactions among peers, such as TalkLife[1] and 7Cups[2]. Unlike in controlled settings, however, responders on these forums often have little or no training in cognitive restructuring. Thus, a natural question to ask is how peer-to-peer conversations on such forums provide support and lead to positive cognitive change. While significant

[1] https://talklife.co/

[2] https://www.7cups.com/

work has studied the content-based [31, 61], affective [21] and supportive [7] parts of these interactions, little is understood about how these factors contribute to cognitive change. A better understanding of these factors can shed light on the underlying processes of cognitive restructuring and help design online forums that provide effective peer support.

In this paper, therefore, we study a dataset of conversations from an online mental health forum, Talklife, and propose methods to identify moments of positive cognitive change. In general, identifying moments of cognitive change is a complex problem because different individuals may express a change in different ways, and improvement is typically detected by questions from validated psychological scales [63]. Thus, while a comprehensive definition of cognitive change from conversational data is impossible, we propose a definition of cognitive change based on change in sentiment that can be operationalized for mental health forums and covers a non-trivial subset of possible cognitive changes. Specifically, we define cognitive change in an online mental health forum as a change in sentiment over any topic that was causing psychiatric distress to an individual. That is, whenever an individual expresses a positive change in sentiment for a topic on which they expressed prior distress, we define it as a **moment of positive cognitive change**, or simply a "moment of change."

Through this restrictive definition, a quantitative analysis of moments of change becomes possible. We construct two ground-truth labels for moments of change: one based on crowdsourced labels of topic and sentiment for each post, and the other based on tracking specific phrases that convey expressions of feeling better or experiencing a change in perspective. The former provides an accurate measure of a moment of change, whereas the latter allows coverage over a larger set of posts from Talklife users.

We design two prediction tasks based on these labels: thread-level and post-level. Given data on all posts from a thread, the goal of a thread-level prediction is to identify if the thread contains a moment of change. This formulation, however, does not allow analysis of a thread as it proceeds. To simulate a prospective analysis, we design a post-level prediction task of deciding whether a given post corresponds to a moment of change, given information on only the prior posts in the same thread. For both prediction tasks, a machine learning model based on linguistic features can identify threads with moments of change with high accuracy (AUC=0.9). Sentiment and affect-based features of the responders' posts are the most predictive of a thread or post having a moment of change. In addition, threads with moments of change are likely to have longer messages and more posts than those without a moment with change.

Throughout, we find the importance of culture-aware models. As predicted by theories of cultural difference [25, 50], we find a marked difference in expression of distress between people from different cultures. In particular, prediction models generalize poorly: when trained on a dataset of Indian users and tested on non-Indian users, the AUC metric for our best performing model drops from 0.90 to 0.68. Threads with moments of change are also more likely to have responders from the same country as the original poster.

While predictive models can be used to track the future potential of a thread leading to a moment of change, they tell us little about *how* cognitive restructuring happens. To do so, we return to our definition of a moment of change and develop a model that explicitly tracks topics and associated sentiment in each post as an approximation for the changes in affect surrounding cognitive restructuring processes. Associating a distressed user's change in sentiment on a topic with specific posts in the forum thread can help to estimate the conversation *pathways* through which cognitive change happens. We call it the "SentiTopic" model. While the theory-driven SentiTopic model has a lower accuracy than the ML-based predictive model, it generalizes well across cultures. Overall, our results demonstrate the value of modeling mental health support conversations, as a way of identifying and directing support to users who may not be getting the help that they need within a thread.

Privacy, Ethics and Disclosure. All data analyzed was sourced (with license and consent) from the TalkLife platform. All personally identifiable information was removed from the data before analysis. In addition, all work was approved by our institution's Institutional Review Board. *This work does not make any treatment recommendations or diagnostic claims.*

2 RELATED WORK

We build upon two streams of past work: computational analysis of cognitive change and well-being in online forums, and aspect-based sentiment analysis and stance detection in social media.

Patterns of cognitive change and support

There is a rich body of work from clinical psychology investigating the processes that lead to positive cognitive, affective, and behavioral change [42]. Using this work as a foundation, recent studies on digital mental health have applied computational techniques to analyze cognitive change.

Drawing from therapeutic discourse analysis [43], Howes et al. [39] and Althoff et al. [6] use computational linguistics to study conversations between people and designated counsellors. Howes et al. find that topic and sentiment features in a conversation are important for predicting change in symptom severity for patients with depression. Instead of characterizing individual conversations, Althoff et al. look at transcripts from each counsellor in aggregate and investigate conversational patterns of a good counsellor. They find that differences in language use and turn-taking distinguish successful counsellors. In this work, we extend these analyses to look conversations in widely participatory online forums with peer supporters, most of whom do not receive any training in counselling. Additionally, unlike Althoff et al.'s exclusive use of crisis-based conversations, online mental health forums ptentially exhibit a broader range of cognitive changes, since suicidal ideation tends to ebb and flow [65] and distress is not exclusively expressed when an individual in crisis [14].

In controlled settings, effectiveness of online peer support for mental health has been demonstrated through randomized controlled trials where participants were explicitly coached to engage with other participants in the forum [38] or identify cognitive distortions when having conversations with others in the peer support community [48]. While this line of work has been effective at reducing symptoms and increasing cognitive reappraisal, it was done within a controlled environment that explicitly encouraged cognitive change, an environment very different than most lightly moderated online mental health forums.

On mental health forums, De Choudhury et al. [23] analyzed specific comments on posts to understand what distinguishes supportive language, with the goal of determining whether a user eventually disclosed that they were suicidal. We use similar features in our work to build models for predicting a moment of change. In addition, while their analysis was focused on users' overall affective language, we consider affect for each topic separately in our proposed SentiTopic model. This is because an increase in overall sentiment may not necessarily indicate a moment of change, but possibly be a sign of avoidance or evasion [53] due to conversational techniques such as pivoting discussion to unrelated topics.

Motivated by the above work, our first research question pertains to predictability of a moment of change.

RQ1: To what extent are moments of change predictable?

RQ1a: Given all posts in a thread, can we predict whether the thread includes a moment of change?

RQ1b: Can we predict whether an individual post will include a moment of change, *before it happens*?

Uncovering pathways that lead to cognitive change

Cognitive change in online forums can also seen from a discursive perspective. In their study of what causes a change in stance in argumentation community on Reddit, Tan et al. [66] used linguistic and metadata-based features to analyze and predict the persuasiveness of an individual post. Features such as the linguistic distance between two posters, the sentiment of the posts, and the complexity of the post were all predictive of whether a particular post caused an original poster to change their mind, a form of cognitive change. Wei et al. similarly predicted the influence of a poster in changing an original poster's mind in a similar context [70].

In mental health forums, however, cognitive change likely occurs through multiple posts that discuss multiple topics, unlike argumentation subcommunities where a specific argument is being debated such as the Change My View subreddit [41] or the Wikipedia editing forum [73]. In addition, the sentiment of a user towards the topic may also be associated with cognitive change, as we discussed above.

From a Natural Language Processing (NLP) perspective, computational linguists have approached the problem of accurately detecting sentiment towards various topics through aspect-based analysis [33]. Applied NLP work in this area has centered around analyzing user reviews [40] or political science data [28]. For instance, Zubiaga et al [74] created a sequential-based model that uses the tree structure of social media text to more accurately model rumor classification in Twitter tasks. A second problem for detecting moments of change is in comparing intensity of sentiment. While researchers have modelled sentiment polarity extensively [55], sentiment intensity is a relatively new field of research [35, 47]. Out of all the current off-the-shelf NLP models, we found VADER [35] the most suitable for estimating strength of sentiment in social media data and thus we use it for our analysis.

In addition to sentiment modeling, we utilize NLP techniques such as word embedding [67] and topic modelling [62] to track changes in sentiment over topics mentioned in a forum thread, leading to our second main research question:
RQ2: What are the conversation pathways that lead to moments of change?

3 DATA

To understand how cognitive change processes happen in an organic online mental health environment, we use data from the peer support forum Talklife [3]. Talklife, a peer support network for mental health founded in 2012 [1], has been designed as a safe space for people to be open about distress and talk through it alongside others [49]. As a result, users of the website discuss diverse topics from normal online banter to self-harm and running away from home [49]. Talklife forum usage is characterized by individuals posting a message, often a question asking for advice or comment expressing distress, and other individuals contributing support in the form of comments on that initial post. Typically, the original poster who started the thread of comments will discuss the issue introduced in their initial post with other responders in the thread. Throughout this paper, we refer to the person who starts a Talklife thread as the *original poster* or *"OP"*, and the people who respond to the thread as *responders* or *"non-OP."* Our dataset consists of a random subsample of all posts on Talklife from May 2012 to June 2018. All posts analyzed were in the Latin script.

Defining the Problem Space

We begin by quantitatively defining a moment of change.

Moment of change*: A positive change in sentiment for the OP on a topic that was mentioned by the OP in their first post, over the course of a conversation in a single forum thread.*

To find threads in which moments of change occur, we construct a sample of threads in which an individual expressed some form of psychiatric distress and filter out the rest. Similar to De Choudhury et al's technique to find expressions of an anxiety diagnosis online [29], we selected a subset of threads that used keywords that are indicators of distress. First, we selected threads that included words labelled by Talklife as *trigger words* including nouns like "suicide" or verbs like "cut." These words are often associated with the expression of distress and thus used on the platform to label content with a trigger warning—an indicator that a post might contain content that may inadvertently evoke upsetting thoughts and harmful behavior [44]. Second, to allow a broader set of threads, we also include adjectives that people with mental illness often use to describe their affective state. Specifically, we used negative adjectives from CBT worksheets [4] that are designed to effectively label one's emotional state. Using this content filtering, we obtain a dataset of 46,832 threads with 415,716 posts.

Deriving a Ground Truth

When using naturalistic social data for behavioral research, creating a ground truth dataset to validate findings can be difficult [24, 72]. Common strategies used to gather ground truth data for mental health and social computing include using domain experts [9, 24] or crowdworkers [20, 68] to manually label and validate data, or using domain knowledge to derive a ground truth [29]. In this work, we utilize crowdwork and domain knowledge to create two different ground-truth datasets.

We first construct ground-truth data directly based on our definition of moment of change. We collected a sample of 2500 posts and designed a crowd-sourcing task to label the sentiment associated with a given post on 7 point Likert scale, with options ranging from strongly negative (-3) to strongly positive (3). Each post was labeled by 3 raters. The average pair-wise difference in rating for the same post was 0.51, indicating that on average, raters disagreed to less than 1 rating point. [3] Moreover, when there was a disagreement, a majority involved a difference of 1 rating point For instance, the post *"I'm depressed and have days where I sob uncontrollably"*, received two strongly negative labels and one negative label. Posts with higher level of disagreement were often those that included both positive and negative content.

We labelled the sentiment of a post as the mean of the sentiments reported by each rater. To identify threads that contained a moment of change, we selected threads in which the mean sentiment increased by at least one rating point. We call this dataset the *annotation-based dataset.*

While the above dataset corresponds to the sentiment-based definition of moment of change, it is limited in size due to the dependence on manual labeling. To create a larger scale dataset, we collected a set of phrases that are associated with the OP feeling better, based on a qualitative analysis of Talklife forum threads. We then did an exhaustive search on our Talklife dataset to select posts that mention any of these phrases and marked them as having moments of change. Specifically, we used regular expression-based phrases to detect when the OP said that they felt better (such as *"I feel much better now"*), or when the OP acknowledged advice from someone else in the thread (such as *"You have a point"* or *"I had never thought of that"*). We omitted any threads in which this happened in the first post after the original post. [4] We refer to this dataset as *pattern-based.*

To validate our larger, pattern-based ground truth labels, we trained a gradient boosting classifier to identify whether a thread has a moment of change, as described in our predictive analysis section. When trained on the pattern-based labels, and tested on the crowdsourced labels, this classifier reported AUC of 0.8, showing consistency between our crowdsourced and pattern-based methods of identifying a ground truth of which threads contain a moment of change.

Splitting up into culture-specific datasets

Since expressions of online support can widely differ between individuals from different cultures [25], we constructed two culture-specific datasets: one corresponding to threads started by Indians and the other to threads started by people from other countries (who we refer to collectively as *"non-Indians"*). We chose to focus our attention to Indian users of Talklife due to the rich literature on the diversity of expression of mental illness in India [60, 71], and past work on Indian members of online mental health communities [25].

Thus, our final dataset comprises of 25,537 threads from Indians without a moment of change and 295 threads from Indian users with a moment of change. Among threads started by non-Indians, we obtain 14,604 threads without a moment

[3] The Fleiss' Kappa inter-rater reliability score was 0.4437, showing moderate agreement. This statistic, however, underestimates rater agreement by considering each of the seven sentiment classes to be equidistant and independent.

[4] In many cases, this occurred when the OP was responding to a post that was deleted later, or if the OP was adding a comment to provide further clarity on their original situation.

of change and 6,396 threads with a moment of change. Additionally, to compare these culture-specific with the general population, we also construct the combined dataset, which we call the Culture Agnostic dataset.

4 DESCRIPTIVE ANALYSIS USING METADATA

We begin our analysis by describing metadata-based aspects of threads that contain a moment of change and comparing it with those that do not. On average, there were 9 messages (mean (μ) = 8.88, median (ν) = 5.0, standard error (se) = 0.12) in a thread (including the original post), with each response to the original post having an average of 17 words ($\mu = 17.11$, $\nu = 12.67$, $se = 0.08$). However, within threads that had a moment of change, there were 12 messages on average ($\mu = 11.70$, $\nu = 8.0$, $se = 0.26$), with an average of 27 words per response to the original post ($\mu = 26.53$, $\nu = 22.0$, $se = 0.21$). When testing the average number of messages in a thread and average number of words in a message between threads with and without moments of change via Welch's t-test, the p-values were less than $2.08x10^{-29}$. This indicates that threads with moments of change have a higher amount of interaction than threads without moments of change, potentially as a result of responders doing more work to express empathy and dissect the cognitive distortions associated with the specific issue that an OP raised in the thread. Moments of change, on average, would happen in the 7th response to the OP's first post ($\mu = 6.92$, $\nu = 4.0$, $se = 0.17$).

These numbers varied when solely looking at the Indian population of Talklife users. On average, in threads initialized by Indian users, there were 6 responses in the thread ($\mu = 6.32$, $\nu = 4.0$, $se = .13$) and an average of 13 words per response ($\mu = 13.34$, $\nu = 9.5$, $se = .08$). However, in threads initialized by Indian users that contained a moment of change, on average, there were 15 responses in the thread ($\mu = 14.60$, $\nu = 9.0$, $se = 1.05$), with an average of 22 words per response ($\mu = 21.64$, $\nu = 17.42$, $se = 0.93$). These differed with a p value of $3.997x10^{-14}$. Moments of change, on average, would happen in the 8th response to the OP's first post ($\mu = 8.14$, $\nu = 5.0$, $se = .60$), which is statistically different from the overall moments of change dataset with p-value of .04. Overall, these results indicate that threads with a moment of change tend to have high amounts of interaction, wherein the relative difference is higher when the OP is Indian.

Since social support on Talklife often transcends national borders, we also looked at the percentage of responders that are from the same country as the OP. For the 93 percent of threads for which location data was available, we found that on average, 41 percent ($\mu = 40.70$, $\nu = 40.0$, $se = 0.15$) of responding posters on the thread were from the same country as the OP. However, in cases where there was a moment of change, an average of 48 percent ($\mu = 47.7$, $\nu = 47.1$, $se = .003$) of responders were from the same country as the OP. This difference is statistically significant (p-value = $4.30x10^{-97}$). In cases where the OP was Indian, this number increased: an average of 54 percent of posters ($\mu = 53.93$, $\nu = 50.0$, $se = 1.14$) were from the same country as the OP, with a statistically significant difference between Indian threads with and without moments of change (p-value = $5.67x10^{-8}$). Thus, across Indian and non-Indian datasets, we find that threads with moments of change are likely to have a higher number of responders from the same country as the OP, in line with past research suggesting that therapists with similar identities as their clients are more positively received [16].

5 FEATURES FOR PREDICTIVE ANALYSIS

Informed by the patterns observed in our descriptive analysis and previous work done in computational psychiatry and computational linguistics, we now investigate whether it is possible to use affective and linguistic features to automatically identify threads that have a moment of change (**RQ1a**), and whether it would be possible to use a similar technique to predict whether a given post would have a moment of change (**RQ1b**).

We divide our features into four main categories:

(1) **LIWC-based**: Features derived from the Linguistic Inquiry and Word Count [58], a psycholinguistic text analysis tool that has been validated for predictive tasks in a variety of mental health related research contexts [17, 22]. We used the 2015 version of LIWC [57] for our analysis.
(2) **Punctuation-based**: Features derived from the punctuation used in the forum, such as exclamation points and question marks.
(3) **Metadata-based**: Features derived from information that does not require any analysis of the content of the text of the posts on the forum.
(4) **Mental health language-based**: Features derived from the occurrence of exact phrases typically used in mental health settings.

LIWC-based Features

Following Gilbert et al. [36] and De Choudhury et al [23], we measured the positive and negative sentiment associated with the words in a post, as well as counts of anger, swear words, and intimacy language. Following Sharma et al. [64], we also implemented a linguistic style matching [51] feature as a proxy for the overall trust and cohesiveness among the people in a thread [37, 64], a core part of the forms of social support that have a positive impact on mental health [46].

Punctuation-based Features

Following Zubiaga et al's sequential method for detecting rumors on Twitter [74], we counted the number of punctuation-level features used in a thread and post, such as the percentage of punctuation that consists of exclamation points, question marks, and periods. We use these features as indicators of questions and exclamations that often occur during cognitive restructuring [54].

Metadata-based Features

Based on our observation that threads with moments of change tend to have higher interaction than other threads, we added the number of posts in a thread and the length of each post as features to our model. Additionally, following Gilbert et al. [36] use of variables indicating reciprocity to predict tie strength, we used the ratio between sent and received messages as a feature, but for our analysis, we took sent messages to mean messages in the thread by the OP, and received messages to mean responses in the thread that were not from the OP. We also included three features based on our descriptive analysis of the relationship between location and moments of change: total number of countries represented within the posters in the thread, the number of posters from the same country as the original poster, and the number of posters with the same gender as the original poster.

Mental Health Language-based Features

Following the use of data from Reddit mental health communities for feature construction in previous work [21, 56], we created a list of the 250 most popular trigrams and four-grams from the Anxiety, Depression, and Suicide Watch Reddit communities in 2015 [10], and counted occurrences of these common support phrases within each post and the overall thread. The goal of this feature is to capture phrases that responders might use when engaging in a cognitive restructuring process with an OP. Additionally, following the use of antidepressant-related language in predicting depression via public Twitter data [22], we also counted mentions of names of medication, sourced from the Wikipedia article that lists all psychiatric medications [2].

6 RESULTS FROM PREDICTIVE MODELS

We now build models for predicting whether a thread or an individual post contains a moment of change, based on our pattern-based ground-truth labels. For the culturally agnostic (CA) data, we construct a dataset with 6410 threads in the train set, 713 validation threads, and 791 threads in the test set. For the culture-specific datasets—Indian and non-Indian—we train on 475 threads each with a validation set of 53 threads and a test set of 62 threads. We fixed the size of the culture-specific datasets to the lowest of the Indian and non-Indian datasets, to allow an equal comparison of models built using them. Throughout, we use a 50-50% split between threads with or without a moment of change.

For all three datasets, we evaluated four different machine learning algorithms. On our validation sets, gradient boosting methods (XGBoost [5]) performed highest out of Random Forest, Support Vector Machine, and Naive Bayes models. Therefore, we report results of the XGBoost model in this paper, using the Area under the Curve (AUC) metric of the Receiver Operating Characteristic (ROC) for each model on the held-out test set: an AUC of 0.5 corresponds to a random classifier and an AUC of 1 corresponds to a perfect classifier.

Thread-level Results

For the thread-level task (**RQ1a**), we build features for each thread and consider a classification task of determining whether a given thread contains a moment of change. We consider three models based on the information that they use: *OP-Only* uses only OP's posts, *Non-OP-only* uses posts from responders, and *All* uses all posts in a thread.

We first report the AUC score for models trained and tested on the same demographic. As seen in Table 1, it is possible to determine whether a thread contains a moment of change, with an AUC score of 0.88 for the culturally agnostic dataset with all features. While LIWC acts as a good approximation for moments of change under the *Non-OP-only* model, it performs worse under the *OP-only* model. When considering only the posts from the OP, metadata features gain importance, as seen by a bigger increase in AUC score with the addition of metadata features in both CA and Indian datasets. Collectively, these results suggest that the language used by responders (as represented by LIWC features) plays the most important part in leading to a moment of change. Without access to responders' posts, metadata features, such as location of the OP or number of words from the OP, become important to detect a moment of change. These conclusions also hold when restricting our analysis to only the Indian dataset, as shown in the bottom panel of Table 1.

Table 2 provides a more granular view of the importance of difference features for prediction. We measure importance through a normalized score of the number of times a feature is used to split data across all trees in the ensemble model. We find that LIWC-based features associated with sentiment, affect and intimacy language are the most predictive. Mentions of n-grams associated with mental health language are also predictive. In comparison, metadata and punctuation features contribute less towards prediction.

To look at cross-cultural differences in characteristics of threads with a moment of change, we design a classification task where we train a model on the Indian dataset and test on the non-Indian test set, and vice-versa. This has the effect of measuring to what extent predictors of moments of change

	LIWC	**LIWC + Punctuation**	**LIWC + Punctuation + Metadata**	**LIWC + Punctuation + Metadata + Language**
CA Dataset	0.87	0.87	0.88	0.88
CA Dataset, only non-OP posts	0.86	0.85	0.85	0.86
CA Dataset, only OP posts	0.68	0.69	0.81	0.81
Indian Dataset	0.89	0.9	0.9	0.9
Indian Dataset, only non-OP posts	0.97	0.97	0.98	0.98
Indian Dataset, only OP posts	0.78	0.73	0.92	0.93

Table 1: Thread-level AUC for models trained on the Culturally Agnostic and the Indian dataset. For both, we obtain high AUC scores (> 0.8) for predicting a moment of change. Models trained on non-OP LIWC features perform better than those trained on OP-only features, suggesting that responders' language plays an important role in detecting a moment of change.

Features	**Culture Agnostic Feature Importances (Thread)**	**Indian Feature Importances (Thread)**	**Culture Agnostic Feature Importances (Post)**	**Indian Feature Importances (Post)**
LIWC Features (Per post in sequence)	**0.497**	**0.448**	**0.525**	**0.516**
Positive Sentiment	0.138	0.136	0.136	0.15
Negative Sentiment	0.114	0.081	0.123	0.124
Affect	0.158	0.176	0.169	0.162
Anger Words	0.053	0.035	0.064	0.063
Swear Words	0.034	0.02	0.033	0.017
LIWC Features (Average over all posts in sequence)	**0.26**	**0.311**	**0.285**	**0.286**
Positive Sentiment	0.028	0.029	0.022	0.023
Negative Sentiment	0.02	0.036	0.022	0.024
Affect	0.013	0.024	0.023	0.025
Anger Words	0	0.006	0.007	0.016
Swear Words	0	0	0.004	0.003
Intimacy Language	0.134	0.15	0.175	0.157
Linguistic Style Matching	0.065	0.066	0.032	0.038
Punctuation Features	**0.034**	**0.049**	**0.065**	**0.05**
Exclamation Points	0.006	0.019	0.026	0
Question Marks	0.012	0.007	0.028	0.038
Period	0.016	0.023	0.011	0.012
Metadata Features	**0.045**	**0.062**	**0.035**	**0.086**
Number of messages in a thread/post	0.024	0.026	0.024	0.04
Average number of words in a post	0	0	0	0
Ratio between sent and received messages	0.005	0.005	0	0.032
Average length of given post	0.015	0.031	0.011	0.014
Total number of countries represented in thread	0	0	0	0
Total number of posters from country of the original poster	0	0	0	0
Total number of posters with same gender as original poster	0.001	0	0	0
Mental Health Language Features	**0.161**	**0.131**	**0.089**	**0.063**
Number of mental health n-grams, derived from Reddit	0.156	0.131	0.07	0.063
Number of medication words used	0.005	0	0.019	0

Table 2: Importance of features used in thread-level and post-level predictive models as measured by proportion of times each feature is used to split the data across all trees in XGBoost. Across all settings, LIWC features are the most predictive.

	Trained on Indian train set	**Trained on non-Indian train set**
Tested on Indian test set	0.90	0.74
Tested on non-Indian test set	0.68	0.86

Table 3: AUC of thread-level models in non-Indian and Indian datasets. Cross- training and testing across cultures results in a significant drop in prediction accuracy.

can be transferred between cultures. We report results from the best performing model above that uses features from all posts. From Table 3, we see that the AUC score when cross-training drops substantially compared to Table 1, from nearly 0.9 when testing on Indian dataset to 0.68 when testing on the non-Indian dataset, for a model fitted on the Indian train dataset. Similar results are obtained for a model fitted using the non-Indian train set. These results signify the importance of culture and the lack of universality of markers of moments of change.

Post-level Results

From thread-level prediction, we turn to the question of whether the next post in a thread includes a moment of change (**RQ1b**). Specifically, given a thread and all posts up to the X*th* post, the task is to predict whether the OP will express a moment of change in the next *(X+1)* post. As before for the thread-level analysis, we consider three data subsets: *OP-only*, *Non-OP-only*, and *All*.

As seen in Table 4, a prediction model with all features obtains AUC scores of more than 0.9 in all three data subsets. When restricting the analysis to an *Indian-only* model, we obtain a similar AUC. While these results are aggregated for predicting a moment of change at any X*th* post, we also study how prediction AUC varies with different X. Setting $X = \{2, 4, 6, 8\}$ leads to an AUC of 0.85, 0.86, 0.91, and 0.9 respectively, showing how increased contextual information from a thread leads to a higher prediction accuracy. Among the different features, LIWC-based features are the most predictive, except for the *OP-only* dataset. As in the thread-level analysis, this indicates that the language used by the responders are the most associated with a moment of change.

	LIWC	LIWC + Punctuation	LIWC + Punctuation + Metadata	LIWC + Punctuation + Metadata + Language
CA Dataset	0.91	0.915	0.9	0.92
CA Dataset, only non-OP posts	0.92	0.926	0.923	0.92
CA Dataset, only OP posts	0.7	0.72	0.91	0.93
Indian Dataset	0.92	0.92	0.947	0.91
Indian Dataset, only non-OP posts	0.92	0.92	0.929	0.947
Indian Dataset, only OP posts	0.669	0.74	0.9645	0.927

Table 4: Post-level AUC for models trained on the Culturally Agnostic and the Indian dataset. We obtain higher AUC scores (> 0.9) than the thread-level models.

Test set	Trained on Indian train set	Trained on non-Indian train dataset
Tested on Indian test set	0.91	0.55
Tested on non-Indian test set	0.59	0.93

Table 5: AUC of post-level models in non-Indian and Indian datasets. As for thread-level models, cross-training and testing across cultures leads to substantial drop in accuracy.

From our feature importance analysis in Table 2, the most important features across both CA and Indian models are LIWC-based features on intimacy, affect, and positive and negative sentiment.

Next, we perform cross-training between non-Indian and Indian datasets for detecting moments of change in a post. We report results for the best performing model from Table 4. As we saw for the thread-level models, cross-training yields substantially lower AUC scores (< 0.6) than training and testing on the same population. This shows that features optimized for a certain demographic do not translate to another, which is consistent with previous literature on the importance of culture in expression and support for mental distress [25, 50].

Overall, our models show that moments of change are predictable and simple features such as LIWC can be used to detect them. We also saw the importance of culturally-aware models. We present related design implications for mental health forums in the Discussion section.

7 GOING BEYOND LIWC: SENTITOPIC

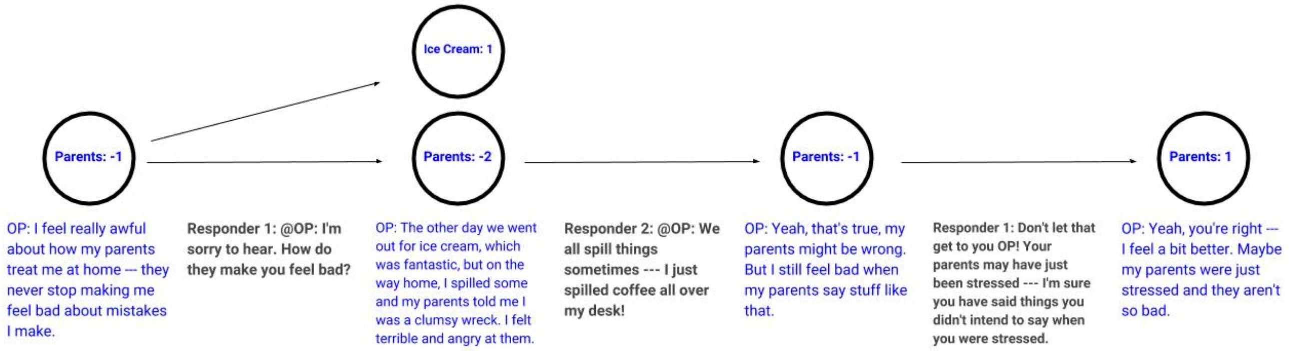
While we were able to detect moments of change in post and thread-level data with reasonable accuracy, predictive results do not tell us much about *how* moments of change happen over the course of a conversation in a thread. In this section, therefore, we construct a model that is derived directly from our definition of a moment of change, rather than the often difficult-to-interpret machine learning models that are commonly used for similar tasks. Intuitively, we aim to build a "SentiTopic" model that extracts topics that caused distress for an original poster, and track the sentiment of the OP's expression towards those topics throughout the thread, as seen in Figure 1. By tracing a moment of change back via the posts in a thread, we can track the pathways of cognitive change in a given thread. Below we present a method based on topic and sentiment analysis that aims to develop this understanding and apply it to the problem of detecting moments of change.

The SentiTopic Model

Given a thread, the SentiTopic model estimates topics and associated sentiment for each post in the thread. To do so, we extract topics from the full thread, then assign a topic to each sentence of a post, and estimate the sentiment associated with that sentence using VADER sentiment analysis [35].

Extract topics in a thread. First, we tackle the problem of extracting topics from each post. We utilize linguistic part-of-speech disambiguation as a preprocessing step and extract all explicit nouns from each post using the Python Natural Language Toolkit [15]. Restricting conversational text to nouns only allows us to focus on the topics or themes that a person mentions in a post. Given the set of nouns in a post, we use a pre-trained model, Sense2Vec [67], to embed each noun as a vector in a common space. We use Sense2Vec because the model accounts for the parts-of-speech *sense* in which words are used. To combine these individual nouns into topics, we use an iterative clustering algorithm that combines nouns into clusters until a similarity criterion is reached. Specifically, we create initial cluster sets such that for each noun n_j in the noun set, we create a cluster of the k most similar nouns to n_j, as measured by cosine similarity. The result of this clustering is m number of k-sized clusters.

We then increase the quality of these clusters by repeating two procedures. First, we merge clusters whose average similarity—as defined by the mean pair-wise similarity between all noun elements—is less than a pre-specified similarity threshold. We then examine each cluster individually, and eliminate noun elements whose average distance with others within the cluster is higher than the same similarity threshold.



OP: I feel really awful about how my parents treat me at home — they never stop making me feel bad about mistakes I make.

Responder 1: @OP: I'm sorry to hear. How do they make you feel bad?

OP: The other day we went out for ice cream, which was fantastic, but on the way home, I spilled some and my parents told me I was a clumsy wreck. I felt terrible and angry at them.

Responder 2: @OP: We all spill things sometimes --- I just spilled coffee all over my desk!

OP: Yeah, that's true, my parents might be wrong. But I still feel bad when my parents say stuff like that.

Responder 1: Don't let that get to you OP! Your parents may have just been stressed --- I'm sure you have said things you didn't intend to say when you were stressed.

OP: Yeah, you're right — I feel a bit better. Maybe my parents were just stressed and they aren't so bad.

Figure 1: A simulated example of how the sentiments around individual topics can change over the course of a conversation in a thread. Each node is labelled as "Topic:Sentiment" corresponding to each post in the thread.

Algorithm 1 Extract-Topics Algorithm

```
1: For every noun $n_j$ in each post, $\phi_j \leftarrow Sense2Vec(n_j), n_j \in Nouns$
2: Create clusters of k-nearest nouns for each distinct noun.
3: Repeat until convergence:
       •Merge similar clusters (avg. similarity < $\tau$)
       •Remove dissimilar words within each cluster (avg. similarity > $\tau$)
```

The combination of the procedures above is the Extract-Topics algorithm (see Algorithm 1). To set the similarity threshold τ, we computed the average cosine distance between words in a synonym dataset generated from NLTK's WordNet implementation [32], which we found to be 0.42. We chose k=5.

Assign topics to each post. To apply the SentiTopic model to the problem of predicting if the next post p_j will contain a moment of change given posts $p_1, p_2 ..., p_{j-1}$, we first feed the entire thread into the model, extracting from it topic set T. Then, we match each sentence from each post in $p_1, p_2 ..., p_{j-1}$ to a topic. Finally, assuming that contiguous sentences are likely to address the same topic, we assign any unassigned sentences the topic of its neighbors.

Estimate sentiment for each topic in a post Finally, given topic labels for each post, we use the VADER tool to estimate sentiment intensity for each post, separately for each topic. Thus, the output of the SentiTopic model is a list of topics present in each post of a thread, and the progression of the sentiment of those topics through the posts.

To benchmark the SentiTopic model in its effectiveness in extracting topics, we tested it against Latent Dirichlet Allocation, a popular technique for topical analysis in clinical psychology [62]. We used a text summarization implementation for LDA [26] to output a set of topics for each post. To evaluate the quality of predicted topic clusters, we tested LDA and SentiTopic against ground truth topic labels, constructed manually by extracting labels for all nouns from a given post and merging similar nouns. We find that the SentiTopic approximates the ground truth topics better: precision and recall for the LDA model is 0.26 and 0.072 respectively, while the SentiTopic model yields 0.82 precision and 0.12 recall. Still, recall for the SentiTopic model is low, possibly because the model discards topic clusters with a single word; thus standalone topics from posts would not be identified.

Testing on Moments of Change

To detect moments of change in posts, we used two specific features derived from the SentiTopic model and report results for the post-level model. The first feature is a binary feature corresponding to whether the post includes a positive shift in sentiment towards at least one topic. The second feature is a vector consisting of the difference in sentiment between the first and the most recent post by the OP for each of the detected topics in a thread.

We test the model on both pattern-based and crowdsourced ground-truth labels, and across the three culture-specific datasets. As shown in Table 6, SentiTopic detects moments of change with nearly 0.7 AUC, which is lower than the predictive models in Section 6. However, since SentiTopic's features are directly derived from our definition of a moment of change, it yields a simpler and more interpretable model. Further, it generalizes better across cultures. Under cross-training, we obtain similar AUC scores (0.67 and 0.66 respectively) when training on the Indian dataset and testing on the non-Indian dataset, and vice-versa. Higher generalizability is probably because SentiTopic focuses directly on

Dataset	Culturally Agnostic Subset	Indian-only Subset	non-Indian only Subset
Pattern-based	0.68	0.7	0.72
Crowd-sourced	0.72	0.69	0.76

Table 6: AUC of post-level SentiTopic model on different datasets. While SentiTopic's prediction accuracy is lower than the ML-based prediction models, it uses a simpler feature set and is potentially more interpretable.

changes in relevant topics and sentiment; however, we leave further evaluation to future work.

Limitations of SentiTopic Model

We consider the SentiTopic model as a first step towards more robust predictive models and briefly point out avenues for improvement.

Lack of Granularity of Sentiment. The SentiTopic model relies heavily on the accuracy of estimated sentiment intensity. Small errors in sentiment can lead to incorrect decisions in detecting sentiment changes between posts. Evaluating VADER's sentiment scores against our crowd-sourced annotated sentiment, we find that root mean squared error of VADER is 1.6, a substantially high error on a [-3,3] scale. Developing an improved sentiment intensity estimator can help in improving the SentiTopic model.

Lack of Disambiguation in Embedding. We used Sense2Vec for our embeddings, which does not disambiguate the meanings of nouns, such as the use of the word "bank" to denote a side of a river as well as a financial institution. Further work may use the ELMO embedding [59] or a customized embedding for mental health language to cluster topics more accurately.

Sentence Representation Limitations. Our model extracts only explicit direct nouns and does not account for direct object pronouns and indirect objects for creating topics, which could potentially improve topic coverage.

8 DISCUSSION

Unlike controlled interventions [38, 48], online mental health forums provide support through organic conversations with a diverse set of participants. In this paper, we analyzed one such online forum, Talklife, to understand patterns of support-giving and how peer conversations can lead to cognitive change for individuals. Among the threads that exhibited psychiatric distress, we found evidence of positive moments of change. A majority of them, however, do not lead to a moment of change, which indicates the potential for improving level of support in online forums. Because these forums provide large-scale data of support conversations, studying them can also shed light on cognitive restructuring processes.

To that end, we proposed a quantitative definition of a moment of change, operationalizing a psychological concept to a definition that can be utilized for online support conversations. Based on this definition, we studied two research questions using a dataset of conversation thread on the online platform. We found that moments of change are predictable, both at a thread-level and at a post-level. Using simple features, such as LIWC-based text analysis features, we were able to predict whether a thread contains a moment of change. When we switched the task to predicting whether a given post has a moment of change, the same features were equally predictive, even without access to any data about the particular post. Our results for **RQ1** indicate that it can be possible to predict whether a thread will lead to a moment of change, even before it actually happens. That said, our results are on a favorable 50-50 sample where there are an equal number of posts with or without moments of change; actual forum data will have a skewed distribution where moments of change are not as common.

In order to understand how moments of change happen, we built the SentiTopic model that explicitly traces change in sentiment over different topics, as a thread progresses. The topic-sentiment summary produced by the model provides an intuitive understanding of a distressed individual's trajectory and can be useful for understanding how conversations lead to cognitive change. While the SentiTopic model makes progress by outlining pathways for a moment of change and is also reasonably predictive of a moment of change, we believe there is a lot more to be done for **RQ2** on *understanding* conversational pathways.

Throughout, we found that culture matters. By dividing the dataset into threads started by Indians and by others, we found differences in how people express distress and in their patterns of support. Strikingly, a predictive model trained on Indians seeking support suffers a significant drop in accuracy when tested on non-Indians, and vice-versa. The effect of culture is lower for prediction results from the SentiTopic model, suggesting that the SentiTopic model might be capturing more stable patterns of cognitive change.

Design Implications

While our results can be broadly used towards enabling technology-assisted interventions for online forums, we describe four major design implications that immediately follow from our analysis.

Routing peer attention. Our post-level prediction results show that prediction of a moment of change is possible, even as a thread conversation is active. Given a thread that shows a low predictive probability of leading to a future post with a moment of change, forum administrators can automatically route new peers to the thread to increase chances of a cognitive change for the original poster. Systems that route attention have been proposed in voluntary work contexts such as Wikipedia [18]; we believe that a similar system can also be used to direct peers' attention to threads that need them. In practice, our model can be fruitfully combined with simpler signals such as number of replies to an original poster.

Connecting peer-based and professional support. Still, peer support is not always perfect, and does not work for all kinds of mental distress. Our thread-level model can be used to identify conversations in the recent past that did not lead to a moment of change. Such threads can be directed to trained counsellors or clinical psychologists who may be better equipped to help the original poster, either on the forum or through special counselling sessions.

Personalized training for peer support. Our models can also be turned towards the peer responders. For instance, by looking at threads with or without moments of change in the recent past, we can identify peer responders who often participate in threads that do not lead to a moment of change. Online forums can design interventions so that these responders are provided personalized tutorials or short trainings whenever they login next, based on the topics and threads they have responded to.

Cross-cultural implications for predictive systems. Finally, we found significant cultural differences among Indians and non-Indians, which likely also transfer to other countries and cultures. When designing predictive systems on mental health forums, we therefore suggest to explicitly account for differences in culture, or test separate models with cross-training before deploying a single one, as we did in Section 6.

Limitations and Future Work

Our work has four key limitations. First, cognitive change is a broad psychological concept. While our definition of moment of change is easy to operationalize, we believe that it covers only a specific part of the different kinds of cognitive change exhibited by individuals. We limited our pattern-based search to moments in which a poster matched specific regular expressions. It is likely, however, that people may feel better without explicitly using one of these phrases, and that there could be other (potentially culturally bound) expressions that signify a moment of change. This suggests the necessity of a deeper qualitative analysis of forum threads to identify and include these diverse phrases and repeat our predictive task. Moreover, while we focused on identifying positive cognitive change on mental health forums, negative cognitive change (in which individuals feel worse, e.g., "thanks, I feel worse now") is certainly possible, and the study of what differentiates those who engage positively and negatively with mental health forums is a potential future direction.

Second, we emphasized that cognitive change is associated with a change in sentiment on a topic that the original poster was distressed about. However, identifying such a topic is non-trivial and therefore we used proxies based on sentiment or self-expressed language, which may not necessarily capture the relevant topic of distress. In addition, we restricted our attention to content that was in English and in the Latin alphabet. While TalkLife is a primarily English speaking platform, it is possible that the effects of culture are much stronger when distress is discussed in different languages. Third, we constructed a favorable 50-50% split for threads with or without moments of change. The actual forum distribution will contain fewer moments of change, and predictive accuracy will likely decrease in that setting.

Finally, while we looked at the importance of simple n-gram features for prediction, our work raises the broader question of what patterns of language lead to a moment of change. This necessitates a deeper study that we leave for future work.

9 CONCLUSION

In this work, we examined cognitive shifts on mental health forums through linguistic, metadata-level, and topic-based tools. We developed ground-truth datasets that allowed us to quantitatively analyze moments of change, and developed predictive models that can detect moments of change with high accuracy. We also built a preliminary model to explicitly track topics and sentiments in a thread. While we obtained reasonable accuracies in detecting moments of change, we believe that our work on computationally analyzing cognitive shifts opens up further questions towards a more granular understanding of *whether* and *how* online peer-to-peer conversations are effective in supporting those who seek help for their distress.

10 ACKNOWLEDGMENTS

We would like to thank Samantha Bernecker for helping us refine our ideas and to provide feedback from a psychological perspective. We thank Talklife for providing us access to their forum data.

REFERENCES

[1] 2013. Australian self harm forum app TalkLife set to shut down. (2013). http://www.abc.net.au/am/content/2013/s3727568.htm

[2] 2018. https://en.wikipedia.org /wiki/List_of_psychiatric_medications_ by_condition_treated.

[3] 2018. https://talklife.co/.

[4] 2018. http://www.hpft.nhs.uk/media/1184/cbt-workshop-booklet_web.pdf.

[5] 2018. XGBoost Documentation. (2018).

[6] 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 (2016), 463.

[7] Nazanin Andalibi, Pinar Öztürk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of# Depression.. In *CSCW*. 1485–1500.

[8] R Michael Bagby, Lena C Quilty, Zindel V Segal, Carolina C McBride, Sidney H Kennedy, and Paul T Costa Jr. 2008. Personality and differential treatment response in major depression: a randomized controlled trial comparing cognitive-behavioural therapy and pharmacotherapy. *The Canadian Journal of Psychiatry* 53, 6 (2008), 361–370.

[9] Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1373–1378.

[10] Jason Baumgartner. 2015. Complete Public Reddit Comments Corpus. (2015). https://archive.org/details/2015_reddit_comments_corpus

[11] Aaron T Beck. 1979. *Cognitive therapy of depression.* Guilford press.

[12] Aaron T Beck. 2008. The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry* 165, 8 (2008), 969–977.

[13] Dror Ben-Zeev, Michael A Young, and Joshua W Madsen. 2009. Retrospective recall of affect in clinically depressed individuals and controls. *Cognition and Emotion* 23, 5 (2009), 1021–1040.

[14] Arpita Bhattacharya, Sean A Munson, Jacob O Wobbrock, and Wanda Pratt. 2017. Design Opportunities for Mental Health Peer Support Technologies.

[15] Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 31.

[16] Raquel R Cabral and Timothy B Smith. 2011. Racial/ethnic matching of clients and therapists in mental health services: A meta-analytic review of preferences, perceptions, and outcomes. *Journal of Counseling Psychology* 58, 4 (2011), 537.

[17] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 51–60.

[18] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 32–41.

[19] Pim Cuijpers, Annemieke Van Straten, and Gerhard Andersson. 2008. Internet-administered cognitive behavior therapy for health problems: a systematic review. *Journal of behavioral medicine* 31, 2 (2008), 169–177.

[20] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 47–56.

[21] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. In *ICWSM*.

[22] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.

[23] Munmun De Choudhury and Emre Kıcıman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, Vol. 2017. NIH Public Access, 32.

[24] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2098–2110.

[25] Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and cross-cultural differences in social media disclosures of mental illness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 353–369.

[26] Guilherme de Oliveira. 2016. Text Summarization. https://github.com/g-deoliveira/TextSummarization.

[27] Robert J DeRubeis, Mark D Evans, Steven D Hollon, Michael J Garvey, William M Grove, and Vicente B Tuason. 1990. How does cognitive therapy work? Cognitive change and symptom change in cognitive therapy and pharmacotherapy for depression. *Journal of Consulting and Clinical Psychology* 58, 6 (1990), 862.

[28] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical Stance Detection for Twitter: A Two-Phase LSTM Model Using Attention. In *European Conference on Information Retrieval*. Springer, 529–536.

[29] Sarmistha Dutta, Jennifer Ma, and Munmun De Choudhury. 2018. Measuring the Impact of Anxiety on Online Social Interactions. (2018).

[30] Ulrich W Ebner-Priemer, Janice Kuo, Stacy Shaw Welch, Tanja Thielgen, Steffen Witte, Martin Bohus, and Marsha M Linehan. 2006. A valence-dependent group-specific recall bias of retrospective self-reports: A study of borderline personality disorder in everyday life. *The Journal of nervous and mental disease* 194, 10 (2006), 774–779.

[31] Marilyn Evans, Lorie Donelle, and Laurie Hume-Loveland. 2012. Social support and online postpartum depression discussion groups: A content analysis. *Patient education and counseling* 87, 3 (2012), 405–410.

[32] Ingo Feinerer and Kurt Hornik. 2017. *wordnet: WordNet Interface.* https://CRAN.R-project.org/package=wordnet R package version 0.1-14.

[33] Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Commun. ACM* 56, 4 (2013), 82–89.

[34] Brandon A Gaudiano. 2008. Cognitive-behavioral therapies: Achievements and challenges. *Evidence-Based Mental Health* 11, 1 (2008), 5.

[35] CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

[36] Eric Gilbert and Karrie Karahalios. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 211–220.

[37] 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 (2010), 3–19.

[38] Kathleen M Griffiths, Andrew J Mackinnon, Dimity A Crisp, Helen Christensen, Kylie Bennett, and Louise Farrer. 2012. The effectiveness of an online support group for members of the community with depression: a randomised controlled trial. *PloS one* 7, 12 (2012), e53244.

[39] Christine Howes, Matthew Purver, and Rose McCabe. 2014. Linguistic indicators of severity and progress in online text-based therapy for depression. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 7–16.

[40] Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 815–824.

[41] Yohan Jo, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn Penstein RosĂŹe, and Graham Neubig. 2018. Attentive Interaction Model: Modeling Changes in View in Argumentation. In *2018 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[42] Alan E Kazdin. 2007. Mediators and mechanisms of change in psychotherapy research. *Annu. Rev. Clin. Psychol.* 3 (2007), 1–27.

[43] William Labov and David Fanshel. 1977. *Therapeutic discourse: Psychotherapy as conversation.* Academic Press.

[44] Stephen P Lewis, Nancy L Heath, Jill M St Denis, and Rick Noble. 2011. The scope of nonsuicidal self-injury on YouTube. *Pediatrics* (2011), peds–2010.

[45] Lorenzo Lorenzo-Luaces, Ramaris E German, and Robert J DeRubeis. 2015. It's complicated: The relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clinical psychology review* 41 (2015), 3–15.

[46] Shery Mead, David Hilton, and Laurie Curtis. 2001. Peer support: A theoretical perspective. *Psychiatric rehabilitation journal* 25, 2 (2001), 134.

[47] Saif M Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. *arXiv preprint arXiv:1708.03696* (2017).

[48] Robert R Morris, Stephen M Schueller, and Rosalind W Picard. 2015. Efficacy of a Web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *Journal of medical Internet research* 17, 3 (2015).

[49] Samantha Murphy. 2015. Uploading depression. *New Scientist* 228, 3046 (2015), 40–43.

[50] Mark Nichter. 1981. Idioms of distress: Alternatives in the expression of psychosocial distress: A case study from South India. *Culture, medicine and psychiatry* 5, 4 (1981), 379–408.

[51] Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21, 4 (2002), 337–360.

[52] Kathleen O'Leary, Stephen M Schueller, Jacob O Wobbrock, and Wanda Pratt. 2018. âĂIJSuddenly, we got to become therapists for each otherâĂİ: Designing Peer Support Chats for Mental Health. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 331.

[53] Akira Otani. 1989. Client resistance in counseling: Its theoretical rationale and taxonomic classification. *Journal of Counseling & Development* 67, 8 (1989), 458–461.

[54] Christine A Padesky. 1993. Socratic questioning: Changing minds or guiding discovery. In *A keynote address delivered at the European Congress of Behavioural and Cognitive Therapies.*

[55] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[56] Sachin R Pendse. 2018. Sochiatrist: Inferring the Relationship Between Emotion and Private Social Messages. (2018). http://cs.brown.edu/research/pubs/theses/masters/2018/pendse.sachin.pdf

[57] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015.* Technical Report.

[58] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001).

[59] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[60] R Raguram, Mitchell G Weiss, Harshad Keval, and SM Channabasavanna. 2001. Cultural dimensions of clinical depression in Bangalore, India. *Anthropology & Medicine* 8, 1 (2001), 31–46.

[61] Stephen A Rains, Emily B Peterson, and Kevin B Wright. 2015. Communicating social support in computer-mediated contexts: A meta-analytic review of content analyses examining support messages shared online among individuals coping with illness. *Communication Monographs* 82, 4 (2015), 403–430.

[62] Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing.* 1348–1353.

[63] Martha Sajatovic and Luis F Ramirez. 2012. *Rating scales in mental health.* JHU Press.

[64] 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.* ACM, 641.

[65] Barbara Stanley and Gregory K Brown. 2012. Safety planning intervention: a brief intervention to mitigate suicide risk. *Cognitive and Behavioral Practice* 19, 2 (2012), 256–264.

[66] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web.* International World Wide Web Conferences Steering Committee, 613–624.

[67] Andrew Trask, Phil Michalak, and John Liu. 2015. sense2vec-A fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388* (2015).

[68] Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* ACM, 74–85.

[69] Sarah Watts, Anna Mackenzie, Cherian Thomas, Al Griskaitis, Louise Mewton, Alishia Williams, and Gavin Andrews. 2013. CBT for depression: a pilot RCT comparing mobile phone vs. computer. *BMC psychiatry* 13, 1 (2013), 49.

[70] Zhongyu Wei, Yang Liu, and Yi Li. 2018. Is This Post Persuasive? Ranking Argumentative Comments in the Online Forum. *Association of Computational Linguistics ACL* (2018).

[71] Mitchell G Weiss, R Raguram, and SM Channabasavanna. 1995. Cultural Dimensions of Psychiatric Diagnosis: A Comparison of DSM–III–R and Illness Explanatory Models in South India. *The British Journal of Psychiatry* 166, 3 (1995), 353–359.

[72] Reza Zafarani and Huan Liu. 2015. Evaluation without ground truth in social media research. *Commun. ACM* 58, 6 (2015), 54–60.

[73] Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. *arXiv preprint arXiv:1805.05345* (2018).

[74] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *arXiv preprint arXiv:1609.09028* (2016).