

# Drivers of Healthy Online Conversations about Loneliness and Depression

by

Lauren Fratamico

B.A., Computer Science, University of California, Berkeley (2013)

M.Sc., Computer Science, University of British Columbia (2016)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

Massachusetts Institute of Technology

June 2019



© Massachusetts Institute of Technology, 2019. All rights reserved

Author

Signature redacted

.....

.....

Program in Media Arts and Sciences

May 24, 2019

Certified by

Signature redacted

.....

.....

Deb Roy

Associate Professor, Program in Media Arts and Sciences

Accepted by

Signature redacted

.....

.....

Tod Machover

Academic Head, Program in Media Arts and Sciences

# **Drivers of Healthy Online Conversations about Loneliness and Depression**

by

Lauren Fratamico

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on May 24,  
2019 in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

## **Abstract:**

Loneliness is becoming a global epidemic. As many as 33% of Americans report being chronically lonely, with similar percentages reported in countries around the world. Additionally, this is a percentage that has risen by as much as 50% in recent years. Many are turning to online forums as a way to connect with others about their feelings of loneliness and to begin to reduce these feelings. However, posts often go unresponded to and online conversations do not take place, perhaps because those conversing did not find a connection between each other, potentially leaving the poster feeling even more lonely. In this thesis, I first define health of conversation for these types of supportive online conversations. I then examine the contributors to conversational health, both in terms of the homophily of the participants and the way in which the participants are conversing. By comparing these characteristics among the spectrum of healthy, supportive, online conversations, I lay the groundwork for being able to facilitate finding optimal conversation partners for those that are feeling lonely. I conclude by envisioning what an interface would look like that would take these factors into account so people can most quickly find the right person to engage with.

Thesis advisor:

Deb Roy

Associate Professor

**Drivers of Healthy Online Conversations about  
Loneliness and Depression**

by

Lauren Fratamico

This thesis has been reviewed and approved by the following committee members (1 of 3):

Rosalind Picard

.....

**Signature redacted**

.....

Professor of Media Arts and Sciences  
MIT Media Lab

**Drivers of Healthy Online Conversations about  
Loneliness and Depression**

by

Lauren Fratamico

This thesis has been reviewed and approved by the following committee members (2 of 3):

  
**Signature redacted**

Iyad Rahwan

.....  
  
Associate Professor of Media Arts and Sciences  
MIT Media Lab

**Drivers of Healthy Online Conversations about  
Loneliness and Depression**

by

Lauren Fratamico

This thesis has been reviewed and approved by the following committee members (3 of 3):

Ethan Zuckerman

.....

Signature redacted

.....

Associate Professor of Media Arts and Sciences

MIT Media Lab

## Acknowledgements

I would like to thank three groups of people. Without all of them, I would not have enjoyed the last 2 years as much as I did.

Firstly, I would like to thank my advisor Deb Roy. I am truly grateful that he gave me the opportunity to study at a place I have been dreaming about since I was a child. I also appreciate all the advice he had given me along the way and the doors he has opened. I'm very excited to be continuing research on conversational health at Twitter after graduation, and it's likely that without having studied the research that I did, I would not have landed that job. I have also very much enjoyed the culture of the lab he created with tons of stimulating people to engage with, mind-opening philosophical conversations, and all of the best labmates I could have imagined.

Secondly, I would like to thank all my friends in Cambridge and at the Media Lab. I'm extremely grateful to have many as both colleagues and friends. I've loved all the late night lab work parties, exercise buddies, and having amazing people to bounce ideas off of about every topic: experimental design, machine learning techniques, visualization, life after graduation, happiness, and many more. Thanks to all of you for keeping me sane these last two years. In particular, I would especially like to thank Arian, Andrea, Bjarke, Cris, Eric, Isabella, Javi, John, Judy, Jules, Marc, Martin, Pranjali, Prashanth, Sanjay, Shayne, Sneha, and Soroush for all the late-night working company, research advice, pep talks, exercise companionship, emotional love, and allowing me to text you at all kinds of hours.

Thirdly, I would like to thank everyone who was interested in hearing about my project and was as astounded as I was about the statistics on loneliness. My intrinsic motivation on this project ebbed and flowed over the two years, but interacting with others who were interested always spiked my motivation and it was phenomenal hearing others speak of the importance of work in this area.

Thanks to everyone for a fantastic past two years of life.

# Contents

<b>1</b>	<b>Introduction</b>	<b>10</b>
<b>2</b>	<b>Related Work</b>	<b>15</b>
2.1	Identifying Healthy Conversations . . . . .	15
2.2	Drivers of Healthy Conversations . . . . .	16
2.2.1	Homophily . . . . .	16
2.2.2	Conversational Style . . . . .	18
2.2.2.1	Sentiment . . . . .	18
2.2.2.2	Discourse Style . . . . .	19
<b>3</b>	<b>Data</b>	<b>20</b>
<b>4</b>	<b>Defining Healthy Conversation</b>	<b>25</b>
4.1	Exploring Traditional Definitions and why they don't work . . . . .	25
4.2	Redefining Health Metrics for (non-toxic and) Supportive Conversations	31
4.2.1	Study Design . . . . .	31
4.2.2	Study and Initial Results . . . . .	33

4.2.3	Defining Conversational Health . . . . .	33
<b>5</b>	<b>Drivers of Healthy Conversation</b>	<b>39</b>
5.1	Feature Engineering . . . . .	39
5.1.1	Metadata about Conversation . . . . .	39
5.1.2	Homophily of Interests . . . . .	40
5.1.3	Conversational Style . . . . .	42
5.2	Results . . . . .	43
5.3	Discussion . . . . .	45
<b>6</b>	<b>Designing an AI System</b>	<b>48</b>
<b>7</b>	<b>Conclusion</b>	<b>52</b>

# Chapter 1

## Introduction

Loneliness is a crippling epidemic around the world. Globally, as much as 40% of people are estimated to experience loneliness at some point in their lives, and this is a percentage that has doubled over the past 50 years. It can impact all ages, ranging from small children to the elderly. Studies show that the reported number of close friends a person has is dropping (from 3 in 1985 to 2 in 2011) [45, 6]. Due to isolation, many people go days without human contact, with an estimated 25% of adults over 75 going a month without seeing another person. Loneliness does not just manifest mentally, but can cause physical damage as well. These physical side-effects have been estimated to cost the US an additional 7 billion dollars in health care costs per year<sup>1</sup>. Other nations are beginning to see that this is a major public health crisis, so much so that the UK recently hired a minister of loneliness. Additionally, this is an epidemic that can impact anyone, regardless of our money, fame, power, beauty, social skills, or personality [14].

---

<sup>1</sup><https://www.thecostofloneliness.org/>

Face-to-face interactions are ideal for combating feelings of isolation, but this is not always possible due to shyness or medically-necessitated bed rest. As a result, many individuals turn to the internet to connect with others. Additionally, when people are lonely, they tend to misinterpret people's faces as more hostile than they are in reality [68, 64, 35], so online interactions may actually be slightly preferable. One place people turn is Reddit, which is full of vibrant, supportive communities where people can converse, offer advice and connect with people they have never met before in person. Research has shown that posting in forms like Reddit about depression can actually improve your mood over time, as indicated by both the language of a user's post becoming more positive [52] and the lexical diversity and readability improving over time [53].

On Reddit, people post about all kinds of topics surrounding loneliness, as can be seen in Figure 1. People like to share stories, ask for advice, share positive updates, and tragically, there's even a whole subreddit on suicidal watch. Fortunately, many of these posts are not going into a vacuum, but are instead being commented on by others in the community, often resulting in vibrant back-and-forth conversation. Some advantages of posting on a website like Reddit include: anonymity/pseudo-anonymity (people have a username on Reddit, but it can be anything and does not link to their real name), forums with specific purposes or topics discussed (so people can post among people that have at least one similar interest), rules and moderators on many forums to guide the types of posts allowed, and high likelihood of the person responding being a human (there are bots on Reddit, but very few).

In general, there are two types of comments to these posts, ones that result in

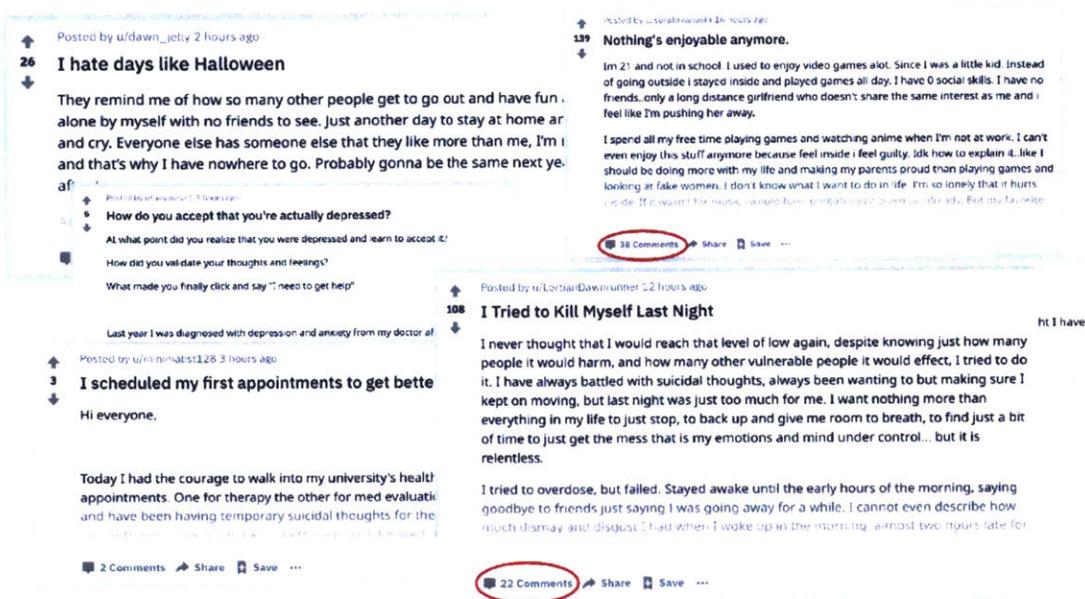


Figure 1: Example posts from Reddit on the topic of loneliness or depression. Post range from shared stories, advice seeking, positive updates, and suicidal posts.

conversation (where two or more people are engaging back and forth), and ones that do not (and instead get no interaction as if the commenter is posting into a void). An example of each of these types can be seen in Figure 2. As face-to-face interactions are the best for combating feelings of isolation, some form of interaction (even if it is just an interaction on an online forum) is better than no interaction.

Additionally, even amongst the conversations that do have interaction, there is a variety to the quality of interaction. Figure 3 shows a selection of some of the conversations that are occurring. As can be seen, there is a variety in the quality of these posts on many dimensions: connection, support, apparent appreciation. In both of the conversations in Figure 3 the conversants are engaged in the conversation, but the conversations take on different forms: in one, advice is offered, and, in the

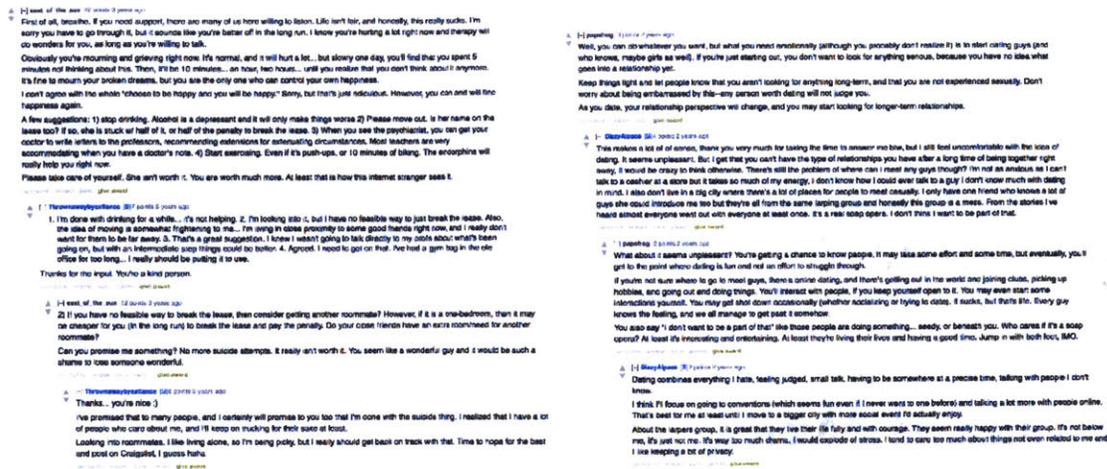


Figure 3: Two examples of supportive conversations on Reddit. The left shows a conversation in which one person is offering advice and the two are engaging, and the right also shows people engaging, but also discussing feelings on an opinion.

account the facilitators of healthy conversation. We would also like to issue a caveat that, given the sensitive nature of some of the Reddit posts, an interface connecting someone to the right online person to talk to may not always be ideal, and instead, some posters may additionally be suggested to a Crisis Hotline, where someone on the other end is more equipped to handle these types of critical conversations.

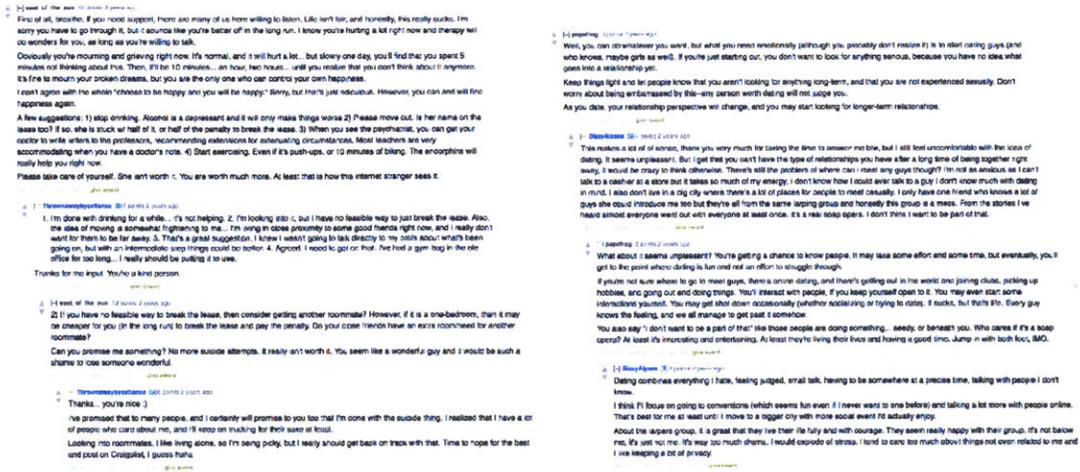


Figure 3: Two examples of supportive conversations on Reddit. The left shows a conversation in which one person is offering advice and the two are engaging, and the right also shows people engaging, but also discussing feelings on an opinion.

account the facilitators of healthy conversation. We would also like to issue a caveat that, given the sensitive nature of some of the Reddit posts, an interface connecting someone to the right online person to talk to may not always be ideal, and instead, some posters may additionally be suggested to a Crisis Hotline, where someone on the other end is more equipped to handle these types of critical conversations.

# Chapter 2

## Related Work

### 2.1 Identifying Healthy Conversations

Given the current climate on the relationship between social media and a variety of negative outcomes (including depression, stress, anxiety, and lack of sleep [2, 8, 27, 30, 40, 67]), many have begun to research how to detect and measure healthy online interactions. Napoles et al. [48, 49] proposed a framework to measure health called ERIC, where comments should be Engaging, Respectful, and/or Informative. They then present a dataset and annotation scheme identifying “good” conversations that occur online along the ERIC framework. Diakopoulos and Naaman also characterized the discourse online, but instead looking at comments made on the Sacramento Bee news site [16]. Others have looked at specific aspects of comment quality (including controversiality [20, 22] and toxicity [23]), or behaviors (such as re-engagement [1] or trolling [10, 46]).

However, with all of these approaches, researchers often overlook both the characteristics of the individuals themselves who are commenting and a broader picture of the linguistic components. This thesis research will combine both behavioral and linguistic features to analyze conversational health.

## 2.2 Drivers of Healthy Conversations

### 2.2.1 Homophily

Homophily is the tendency of individuals to associate and bond with those that are like themselves, or, proverbially, that “birds of a feather flock together”, and those that bond are more likely to have better conversations as they have more to talk about. Network analysis has consistently shown that more homophilous individuals tend to associate with each other, whether people are choosing adolescent friendships [26], romantic partners [28], or even doctors [25]. This is not just the case for in-person interactions, but in the online space as well, with researchers looking at friendships in Myspace [62] and on an online messaging platform [39]. Additionally, Watts et al. [65] found, when analyzing in-person social networks, that individuals even explicitly understand a measure of social distance between themselves and others, with similarity being judged along multiple social dimensions.

Typical definitions of homophily use demographics to define the similarity between people, for example using race, sex, gender, or common language [5, 12, 33, 60]. While these are prominent markers of similarity between two people, they are harder to mine from a website like Reddit where people do not have profiles created for

themselves that display this information as is the case on a website like Facebook. What can be mined from a site like Reddit is the interests of a user, based on what subreddits they're posting on or what they have talked about in the past, and homophily can also be defined based on shared interests (as opposed to shared demographic qualities). It is evident that in-person friend groups tend to form based on these interest-based homophilous characteristics in addition to demographics-based ones [11], but on a forum like Reddit, people are interacting that do not know each other in person. In some cases, communities tend to form around common interests online, even if the people do not know each other. Chang et al. [9] studied activity on Pinterest and found that repinning (resharing) happened more amongst those that were interested in the same topic than those that were previously friends and following each other on Pinterest, showing that shared interests may be a stronger driver of activity than social connections. This indicates that communities do form online among those that have not previously interacted in real-life and that these communities are interest-driven. Communities do form on Reddit based on common interest, as is emphasized by the presence of subreddits (a subreddit being a community, by definition). However, it is unclear if this is the case for Reddit conversations and leaves the question unanswered of if people communicate better on Reddit who don't know each other, but do have a number of interests in common. Ren et al. [56] tested theories about community attachment by forming groups on the MovieLens film recommendation site based on similarity of movie tastes. They found that people felt more attached to those that had similar tastes in movies as they did, and closer in general to the groups where people were grouped based on similarity. This is

perhaps more evidence that those on Reddit that have stronger homophilous bonds would converse more. On the contrary, Bisgin et al. [3] found that interest-based homophily was not enough to construct new friendships on platforms like BlogCatalog, Last.fm, and LiveJournal. However, the Reddit users examined in this thesis do not need to form friendships with each other, but instead just have healthy one-time conversations. This is a much lower bar than forming a friendship. Given all the research on homophily and increased interaction, a measure of homophily will be used in this research to understand its effect on the health of online conversations.

## **2.2.2 Conversational Style**

Another angle of homophily could be similarity of linguistic choices. In in-person conversations, the psychological theory of communication accommodation suggests that participants in conversations tend to converge to the same language and behaviors as their conversational partner [7]. This can even be modeled in online conversations as has been done with Twitter [13, 17]. Linguistic choices can be measured by sentiment and discourse style of posts in a conversation, and are also included in this research to understand its effect on the health of online conversations.

### **2.2.2.1 Sentiment**

Sentiment analysis algorithms are widely used to identify the underlying viewpoint in a span of text by predicting a polarity of sentiment [50] and classify text in terms of its sentiment [51]. Additionally, models have been built for a variety of data sources, with training data coming from sources such as Twitter [19, 24, 58], movie reviews

[61], and product reviews [44]. In the psychology literature, the LIWC [54] is often used to automatically annotate text for its sentiment properties, using theories of emotion such as Ekman’s six basic emotions [18] and Plutchik’s eight basic emotions [55].

#### **2.2.2.2 Discourse Style**

The psychology community has developed a number of scales to manually annotate conversations for different styles. These are often used on recorded in-person conversations, but could be applied to online conversations as well. For each of these scales, the conversations are annotated for both sentiment qualities along with the way in which they are communicating (eg, humor) [57, 66]. Parallel approaches were made to remove the sentiment content from the coding schemes and instead focus on solely the type of the communication. Van Dijk [63] discusses some characteristics of discourse including functionality, meaningfulness, and goal-directedness. Herring has begun to adapt these for use in online and computer mediated conversations [29]. LIWC [54] is also used as a dictionary-based method to label communication methods in writing. The computer science community has also begun to develop analysis methods based on the sociology research in discourse analysis. Zhang et al. [70] labeled and built a classifier for conversations on Reddit for 9 discourse styles: Question, Answer, Announcement, Appreciation, Agreement, Elaboration, Disagreement, Humor, and Negative Reaction.

# Chapter 3

## Data

Reddit<sup>1</sup> is a social news aggregation website where users share posts on a variety of topics, comment on these posts, and up/down vote all submitted content [37, 43]. Posts are self-categorized by their poster into one of thousands of ‘subreddits’. Each subreddit forms a community, and they vary widely in topics, ranging from gaming to fitness to food. Each subreddit is also run like a community with moderators to help those in the community adhere to the community’s rules (eg, no vulgarity) [42]. While the site also features chatrooms where people can have realtime discussions with each other, the bulk of the activity is on the posts made within the subreddits [38].

On July 2, 2015, Jason Baumgartner released a complete copy of Reddit available for public download. This contains over 1.7 billion Reddit posts and their comments, along with all available metadata (author, subreddit, position in comment tree, and

---

<sup>1</sup>reddit.com

other fields that are available through Reddit's API<sup>2</sup>. Since then, the entire dataset has been uploaded to Google's BigQuery<sup>3</sup> and updated so that it now contains a more complete version of Reddit posts [21] and all posts from Reddit's creation (in June 2005) through October 2018. Many researchers have begun to use this Reddit dataset to investigate a wide variety of questions, including examining online hate speech [59] and detecting sarcasm [32], but so far no one has used Reddit to investigate the health of discussions about loneliness.

To compose the dataset for my investigation, I pulled all posts that contained one of the following 6 words and phrases: lonely, loneliness, feel alone, lonesome, depressed, and depression. This totals 2.52 million posts between June 2005 and October 2018. Additionally, I pulled all comments on those posts (over 25.4 million comments) and all previous posts of authors of those posts and comments. Posts come from 50,404 different subreddits. Table 3.1 shows the top 20 represented subreddits (by number of posts). Many of the posts come from mental health or relationship subreddits. Note the relative number of comments per post. r/AskReddit is a subreddit with a much larger following than the others, and, as a result, has a higher average number of comments per post.

As this analysis aimed to focus on discussion about loneliness and depression on Reddit, postprocessing was required to remove threads that were instead sharing images or looking for exchanges. For example r/r4r is an 18+ community to find "platonic or non-platonic friends", and most posts are ones eliciting sexual partners.

---

<sup>2</sup>[https://www.reddit.com/r/datasets/comments/3bxlg7/i\\_have\\_every\\_publicly\\_available\\_reddit\\_comment/](https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/)

<sup>3</sup>[https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit\\_posts](https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_posts) and [https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit\\_comments](https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments)

Table 3.1: Total posts and comments on the top 20 represented subreddits.

SUBREDDIT	TOTAL POSTS	TOTAL COMMENTS
DEPRESSION	632,697	2,783,807
RELATIONSHIPS	84,916	1,259,167
OFFMYCHEST	69,189	262,310
ASKREDDIT	65,218	1,595,738
SUICIDEWATCH	56,250	351,913
NOFAP	49,174	362,914
RELATIONSHIP_ADVICE	35,449	340,811
ADVICE	28,746	142,560
R4R	27,312	76,721
ANXIETY	24,750	122,397
LONELY	22,588	123,109
RAISEDBYNARCISSISTS	22,147	217,887
BIPOLAR	21,453	163,445
DIRTYPENPALS	20,201	9,247
STOPDRINKING	16,611	199,060
MENTALHEALTH	15,261	58,392
DRUGS	14,787	263,540
TREES	14,397	160,828
TEENAGERS	13,912	178,227
NEEDAFRIEND	13,661	55,669

While this is a community where many people are lonely and are looking for others, the conversations tend to be more functional rather than supportive and online. Additionally, roughly 4% of all posts were porn content, where the text of a posting was something along the lines of "I'm so lonely", and then a naked picture was attached. To remove these two types of posts, I pruned entire subreddits instead of pruning individual posts as the false negative rate was too high. To select the subreddits to include in my analysis, I went through the top 100 subreddits by count of posts on loneliness and included them if they contained primarily discussion-type

posts. Of the top 100, this included 55 subreddits that can be seen in Table 3.2. In Table 3.2, I have categorized the subreddits into 5 broad categories:

- Mental Health - includes the subreddits on depression, loneliness, and many of the specific mental health disorders (e.g. r/bipolar and r/ADHD)
- Relationships - includes the subreddits that cover topics of romantic and non-romantic relationships
- Community - includes the subreddits of specific groups of people (e.g. lesbians, pregnant women, moms) as these are places where those groups of people go to discuss many topics including loneliness and depression
- Ending Addiction - includes the subreddits where people discuss overcoming certain addictions (e.g. porn (r/NoFap and r/pornfree), eating (r/loseit), marijuana (r/leaves))
- Other - includes subreddits where people go for general advice and to share general stories

Table 3.2: The 55 subreddits whose posts and comments are included in this analysis, listed in the 5 broad categories they fall into.

SUBREDDIT				
MENTAL HEALTH	RELATIONSHIPS	COMMUNITY	ENDING ADDICTION	OTHER
DEPRESSION	RELATIONSHIPS	ASKTRANSGENDER	NOFAP	OFFMYCHEST
SUICIDEWATCH	RELATIONSHIP_ADVICE	TWOXCHROMOSOMES	STOPDRINKING	ADVICE
ANXIETY	RAISEDBYNARCISSISTS	ASKGAYBROS	LOSEIT	DRUGS
LONELY	FOREVERALONE	ASKMEN	LEAVES	CASUALCONVERSATION
BIPOLAR	BREAKUPS	BABYBUMPS	PORNFREE	MMFB
MENTALHEALTH	DEADBEDROOMS	KETO	NOOTROPICS	NEEDADVICE
DEPRESSION_HELP	DATING_ADVICE	FITNESS	STOPSMOKING	TRUEOFFMYCHEST
ADHD	UNSENTLETTERS	BREAKINGMOM		RANT
BPD	LONGDISTANCE	ACTUALLESBIANS		PERSONALFINANCE
BIPOLARREDDIT	EXNOCONTACT			ASKTRP
SOCIALANXIETY	DIVORCE			JOBS
DEPRESSED				
ASPERGERS				
SELFHARM				
SOCIALSKILLS				
GETTING_OVER_IT				
DEPRESSIONREGIMENS				

# Chapter 4

## Defining Healthy Conversation

### 4.1 Exploring Traditional Definitions and why they don't work

As studied in many other works and mentioned in the related work section, typically healthy conversation is defined as not toxic conversation. As such, this was the initial approach for how to separate healthy and non healthy conversation. Oftentimes this is done with Google's Perspective API<sup>1</sup>.

The Perspective API is the industry standard for how toxicity is measured. It is trained on a dataset of comments on Wikipedia Talk Pages which are the ones where those editing can discuss improvements to Wikipedia pages. Each of these over 160k comments was then annotated by 10 people for how toxic the comment was where a "toxic" comment was defined as one which is a "rude, disrespectful, or unreasonable

---

<sup>1</sup>[www.perspectiveapi.com](http://www.perspectiveapi.com)

comment that is likely to make people leave a discussion". A convolutional neural net (CNN) was then trained on this dataset so that other pieces of text could be classified. Google made their API publicly available for anyone to use, with the main intent to identify harassment on social media or as first-pass at filtering comments on news websites<sup>2</sup>. They further partnered with The New York Times to annotate additional data and include different dimensions of toxicity as part of what their API is able to classify and return. Annotators annotated for many dimensions of toxicity, but the two I will use are among their most common. Presented below are the results for labeling the Reddit comments on three of the toxicity measure defined by the Perspective API. These measures are:

- Toxicity - rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion. Trained on Wikipedia comment data.
- Obscene - Obscene or vulgar language such as cursing. Trained on New York Times data tagged by their content moderation team.
- Inflammatory - Intending to provoke or inflame. Trained on New York Times data tagged by their content moderation team.

Using Google's Perspective API, each of the 25 million comments in my reddit loneliness dataset were labeled with their Toxicity, Obscene, and Inflammatory scores. Each of these scores is a value between 0 and 1, where 1 is the most toxic. The distribution of scores can be seen in Table 4.1 and Figure 4. As can be seen in both the table and the figure, the average toxicity score (for each of the three toxi-

---

<sup>2</sup><https://www.blog.google/technology/ai/new-york-times-using-ai-host-better-conversations/>

city scores) is low, with averages in the 0.2-0.26 range, for each of the three scores. To give an idea of the types of comments across the range of scores, a few selected comments can be found in Table 4.2 with their respective Perspective API toxicity scores. Note how benign the comments are until the very high toxicity scores (Toxicity > .99), and even then, ones can have high toxicity scores and solely contain obscene words. A selection of the most toxic comments are shown in Table 4.3. Note that even though all of the most toxic comments contain obscene words, many are actually supportive of the other person in the conversation (eg, "♥ to you. Fuck those assholes."). In fact, most comments below a score of 0.98 are not negatively contributing to the conversation, unless the other participant is hugely offended by obscene words. Table 4.4 summarizes the percent of comments that have scores above certain thresholds. Because so few comments could actually be labeled as toxic (and therefore unhealthy), we had to explore other ways to label health of conversation for this analysis. Additionally, in the wake of many reports on the toxicity and harms of social media, it is heartwarming to see that there is at least one corner of the internet where supportive, non-toxic conversations are happening.

It should also be considered that because the data that the Perspective API is trained on comes from New York Times moderated comments, a comment could be too toxic to publish on a New York Times article, but not too toxic to be part of a Reddit discussion. This is another reason that we had to further define metrics for conversational health in the informal online space.

Table 4.1: Perspective API Toxicity scores

	TOXICITY	OBSCENE	INFLAMMATORY
COUNT	2.540292E+07	2.540292E+07	2.540292E+07
MEAN	2.069E-01	2.552E-01	2.617E-01
STD	2.266E-01	3.427E-01	2.149E-01
MIN	6.343E-04	3.312E-09	9.915E-09
25%	6.189E-02	2.744E-02	8.827E-02
50%	1.092E-01	7.127E-02	1.867E-01
75%	2.552E-01	3.460E-01	4.247E-01
MAX	9.978E-01	1.000E+00	1.000E+00

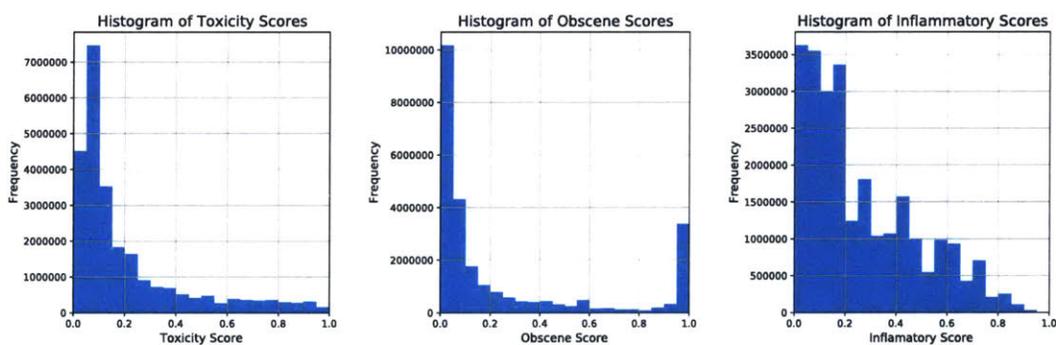


Figure 4: Distribution of the Perspective API scores for Toxicity, Obscene, and Inflammatory. Note the right skew of each.

Table 4.2: A Sample of Reddit Comments with their respective Perspective Toxicity Scores. Sorted by increasing toxicity score.

PERSPECTIVE API SCORE			COMMENT	COMMENT ID
TOXICITY	OBSCENE	INFLAMMATORY		
0.0057	0.0062	0.0146	THANKS SO MUCH. AND YOU'RE RIGHT. THE MORE WE, AS A COMMUNITY CAN SHARE IN TERMS OF EXPERIENCES AND RESOURCES, THE STRONGER WE BECOME.	DJVEAAW
0.0479	0.3372	0.3679	DUNNO IF I CAN START A GAMING CLUB TBH, EGYPTIAN UNI :P	C69LQQW
0.2515	0.9661	0.6995	THERE IS A RISK OF MISSING OUT ON A HELL OF A RIDE YOU THINK THE BORING IS YOUR ALLY, BUT YOU MERELY ADOPTED THE BORING AS A DEFENSE MECHANISM. I WAS BORN IN IT, MOLDED BY IT. MY FIRST WORDS WERE EH.	CLJ15k0
0.3134	0.0198	0.2662	FEEL SORRY FOR YOU? WHY DO YOU THINK THAT? THINK ABOUT HOW YOU COMMUNICATE WITH THEM. DO YOU BRING POSITIVE INTERACTION TO THEM, OR ARE YOU ALWAYS DOWN? YOU CONTROL THEIR PERCEPTION OF YOU. IF YOU ACT MISERABLE, YEAH, THEY ARE GOING TO FEEL SORRY FOR YOU. BUT THAT IS SOMETHING YOU CAN CHANGE.	DWL207J
0.3601	0.6574	0.2132	SOUNDS LIKE YOU IDENTIFIED YOUR PROBLEM. THAT'S A GOOD FIRST STEP.	DZ419QP
0.4985	0.0227	0.7079	YOU AREN'T TRAPPED! MOVE OUT, TRAVEL, PLAN A LONG HIKE, RELOCATE CITIES ENTIRELY, APPRENTICE WITH A TRADESMAN, ETC. IF YOU'VE GOT A COUPLE THOUSAND SAVED UP YOU CAN DO MOST OF THESE THINGS NO PROBELM, AND IF NOT, WELL NOW YOU'VE GOT A GOAL ASSOCIATED WITH YOUR MISERABLE JOB. GOOD LUCK!	C9SBLMN
0.6701	0.6574	0.7236	VIRGINITY FEELS PERMANENT, AND THEN INCONSEQUENTIAL. IF YOU LOSE IT TO A HOOKER.. SO WHAT? YOU'LL LOOK BACK ON IT AND LAUGH AND CRINGE AND MAYBE LIE TO YOUR FRIENDS WHEN THE CONVERSATION TURNS TO THAT LATE IN THE EVENING OR EARLY IN THE MORN'... SAME AS EVERYONE!	CLU8EGX
0.7431	0.9833	0.6872	HANG IN THERE. I WENT THROUGH EXACTLY WHAT YOU ARE AND IT SUCKS. KEEP FIGHTING ON YOUR DISABILITY, IT IS YOUR MONEY. STAY STRONG AND IF YOU NEED SOMEONE TO TALK TO, PM ME. I AM UP ALL HOURS.	C8RL328
0.8341	0.9937	0.5967	CONGRATS BROTHER, KEEP IT REAL! I'M WITH A CRAZY BITCH NOW THIS SUB IS LIKE MY LIGHTHOUSE. IF TWO LITTLE KIDS WEREN'T INVOLVED, HER ASS WOULD BE OUT THE FRONT DOOR. READING MGTOW JUST STRENGTHENS MY SOLVE. BITE THE BULLET LIVE WITH HER SERVE HER AND HOPE SHE FLIPS HER SHIT. SO I CALL THE COPS AND HAVE HER ASS TAKEN AWAY. MAYBE PUT MY INVESTMENTS AND PAY FOR A THERAPIST TO EVALUATE HER AND USE IT IN COURT.	UPNWTP
0.9257	0.9937	0.4205	IT'S NOT CHILDISH TO ME IT'S A SOURCE OF INSECURITY TO YOU AND IT FUCKING SUCKS TO FEEL THAT WAY WANTING TO FEEL SECURE IN YOUR APPEARANCE ISN'T CHILDISH TO ME	E6WX2K3
0.9907	0.9937	0.3599	THIS IS SOME BULLSHIT! DON'T YOU DARE TAKE THAT STUPID BITCH BACK! I AM A WOMAN AND I WOULD NEVER DO THIS TO ANY MAN! FUCKING STUPID! I AM SOOOOO SORRY FOR YOUR SITUATION HOWEVER YOU SEEM LIKE A TOUGH DUDE SO I KNOW YOU'LL BE SMART AND NOT TAKE THAT FUCKING SHIT FROM NOBODY! UP VOTES FOR YOU!!!!	CEUJEG3
0.9930	0.9833	0.4461	YOU'RE A FUCKING MORON.	CBJNODB

Table 4.3: A Sample of the most toxic Reddit Comments with their respective Perspective Toxicity Scores

PERSPECTIVE API SCORE			COMMENT	COMMENT ID
TOXICITY	OBSCENE	INFLAMMATORY		
0.9979	0.9927	0.4459	FUCK YOU YOU STUPID FAT UGLY GAY SACK OF SHIT	CNKVXTD
0.9964	0.9920	0.3149	YOU'RE A FUCKING ASSHOLE OP. FUCK YOU.	C71AMJG
0.9912	0.9920	0.2675	GO FUCK YOURSELF.	CWTH14J
0.9907	0.9937	0.3882	FUCK CANCER!!! YOU ARE BRAVE AS FUCK!!!!!! LOOK IT IN THE EYE AND SAY FUCK YOU CANCER!!!!!!!!!! SORRY LOST MY PARENTS AND GRANDMA TO CANCER AND WHENEVER I READ POSTS LIKE THIS I LOSE MY SHIT. YOU ARE YOUNG AND FROM WHAT I READ YOU'RE TOUGH AS WELL. I WISH NOTHING BUT THE BEST FOR YOU!	C34VKFG
0.9898	0.9923	0.4501	SHUT THE FUCK UP UNFUNNY IDIOT	DVMSJ3L
0.9892	0.9892	0.3639	FUCK OFF	DB00Y79
0.9839	0.9920	0.3667	FUCK THAT GUY.	E4H6DBO
0.9838	0.9937	0.4700	♡ TO YOU. FUCK THOSE ASSHOLES.	DYTCK5I
0.9838	0.9937	0.4385	SOMEONE DOWNVOTED YOU, FUCKING ASSHOLES. NAH, I APPRECIATE YOU. TRULY.	DSYD0JG
0.9837	0.9892	0.5527	FUCK THAT BITCH, KEEP DOING YOU AND CONTINUE KICKING ASS AND GET WHAT YOU ARE PURSUING.	D6N6KOO
0.9729	0.9918	0.3664	FUCK THAT GUY, YOU DESERVE BETTER	COKVGN3
0.9687	0.9910	0.4537	THAT IS A FUCKING AWFUL WORKPLACE	D62AH5Q

Table 4.4: Percent of Perspective API Scores about score thresholds.

SCORE	TOXICITY	OBSCENE	INFLAMMATORY
.95	1.7286%	13.2724%	0.1046%
.90	0.5635%	14.5276%	0.0006%

## 4.2 Redefining Health Metrics for (non-toxic and) Supportive Conversations

As traditional definitions of health of conversation were not going to be sufficient, we had to redefine measures of conversational health. In order to do this, we designed a mechanical turk task to label the health of a conversation and to label certain dimensions of the health. In particular, researchers have coded supportive conversations for ability to acknowledge the other’s viewpoint, engage via followup questions, and listen [31, 34, 36]. The choices of what to ask mechanical turk workers to label health of conversation was based on psychology research on supportive conversations. Final questions can be seen in Figure 6 and are further explained below.

### 4.2.1 Study Design

The study consisted of two parts. The first is shown in Figure 5. Here, we presented two of the Reddit conversations. These were randomly sampled by a stratified random sampling according to the 5 groups that the subreddits were categorized into. Only dyadic conversations were shown, so each conversation just has two people talking back and forth, with the requirement that there must be at least three comments within the conversation (i.e.,  $P1 > P2 > P1$ ). Participants were asked to read through both conversations, as mentioned in the instructions seen at the top of Figure 5 and were told that skimming was okay. They also were prompted that they would have to make

an assessment of which conversation was better and answer a few questions about each conversation. As evaluating health of a conversation is a subjective and challenging task, we chose to have participants view two conversations at a time so that they could compare the conversations and styles of interaction so as to better be able to rate the conversations.

Within the passages were two attention checks put in place to attempt to be able to filter out participants who were not actually taking the task and instead just randomly selecting answers. As can be seen in Figure 5, participants were required to check both of the boxes that were put in the text that specified "Check this box to ensure you are paying attention". The boxes were fairly obvious, so this way, even if a participant was skimming the conversation, they would still see the box. But, if participants were instead just skipping to the questions, they would miss the checkboxes.

The second part of the mechanical turk task is shown in Figure 6. Here, participants were first asked which conversation was better. Then, they were asked a series of questions about each conversation. Four of the questions were Likert scale questions and asked about different dimensions of conversational health: engagement, supportiveness, connection, and appreciation. Initially, the question about appreciation was meant to be used as an attention check question where we could easily check if words like "thanks" or "appreciate" were used and check this against the response to the question about if appreciation was explicitly shown. However, participants interpreted "appreciation" in a much broader way than I had envisioned. For example, it appearing that a suggestion was going to be taken into account by the other person in the conversation was enough to be seen as being appreciative. We then added a short answer question where people could describe in a few words how the conversation participants showed their appreciation. This proved to be valuable to further ensure that mechanical turk workers were paying attention to the task as they had to write something sensible.

There was one additional question on the survey, which was optional, and asked if there were any other reasons why one conversation was better than the other.

Three small pilots were done on this task and minor modifications were made to increase the rate of quality work done by the mechanical turk workers. The final version of the task is what is described above and shown in Figures 5 and 6.

## 4.2.2 Study and Initial Results

To collect the dataset we will use for labeling conversational health, we ran a mechanical turk study with 1000 HITs, and required 3 workers per HIT. Participants were allowed to do more than one HIT, and were paid \$0.25 per HIT they completed. The 1000 HITs were chosen via stratified random sampling of the 5 subreddit categories defined in Table 3.2. As each HIT contained two dyadic conversations, 400 conversations were selected from each category. As a result, we had labels for 1000 pairs of conversations, in terms of which was better than the other, and likert style ratings for 4 dimensions of conversational health for 2000 conversations.

As participants were asked to select which of two conversations was better, we can use that as a baseline value for how often the participants agree. On 61.4% of HITs, participants agreed which was the better conversation. Random chance for this would be 25% as there are three people making binary choices. Given that rating conversations is a challenging and subjective task, we were pleased with the high agreement among raters.

## 4.2.3 Defining Conversational Health

To form a final measure of conversational health, we wanted to combine the 4 subdimensions that we had surveyed people about: support, connection, engagement, and appreciation. To do this, we took each one with equal weight. To ensure that equally

Table 4.5: Health of Conversation Score Agreement with Selection of Better Conversation

	AGREEMENT PERCENTAGE				
	ALL SCORES	W/O ENGAGEMENT	W/O SUPPORT	W/O CONNECTION	W/O APPRECIATION
ALL 3 AGREED	93.76%	93.82%	93.82%	93.82%	93.70%
GENERAL	83.04%	83.29%	83.77%	83.09%	83.92%

weighting the scores made for a good health of conversation score, we computed the percent of times that that score for each conversation agreed with which conversation each person rated as better. Additionally, we looked at the agreement if calculating a score with each of the 4 health dimensions removed. We did this for just conversations where all three people selected the same conversation as better and for all rated conversations. These percentages are summarized in Table 4.5. Since removing none of the dimensions changed the agreement percentage much, we decided to keep all dimensions. The final composite health of conversation score was the sum of the scores for each dimension normalized to a value between 0 and 1, giving a final health of conversation score from 0 to 1. Additionally, the final score used in analysis was the score averaged across all raters. The distribution of scores can be seen in Figure 7. Note the slight left-skew, but overall fairly normal distribution. It should also be noted that this is a measure that should only be used for healthy and supportive conversations as that was the subset of data it was labeled on. Two Reddit postings with their health scores can be seen in Figure 8.

Below is a pair of conversations. Please read the conversations (it is okay to skim) and answer which conversation is better and a few questions about each conversation.

There may be a few attention check questions, so please be sure to pay attention to those if you want to receive full pay.

Thanks! I really appreciate you helping my research on conversational health! -Lauren ;)

*This HIT is part of a MIT scientific research project. Your decision to complete this HIT is voluntary. There is no way for us to identify you. The only information we will have, in addition to your responses, is the time at which you completed the survey. The results of the research may be presented at scientific meetings or published in scientific journals. Clicking on the 'SUBMIT' button on the bottom of this page indicates that you are at least 18 years of age and agree to complete this HIT voluntarily. Please email Lauren Fratamico at fratamico@mit.edu with any questions or concerns.*

#### Conversation 1

**P1:** I just wanted to say I'm really sorry to hear about your husband and the felony charges, I've never been married so I don't know how that feels but I was accused of sexual assault (which I didn't do) when I was 16 and that was I first got depressed (it really ruined my reputation and when I'm visiting home and have to see or interact with all the people who were involved my depression gets worse). However I'm in the same place with regards to suicidal thoughts, I'm in my home town for the summer and because of what happened back then I'm alone and have a lot of trouble meeting new people. I also was recently diagnosed with a chronic illness that WAS curable if doctors had caught it earlier, but now I'm essentially going to feel nauseous my entire life and I'm really fucking scared. Like really fucking scared, and even more so because lately I've been thinking about what the point of even living is if my quality of life is going to be so much worse. My family keep trying to tell me God is trying to tell me something or get me to change, but they don't know how it feels and how saying that makes it worse. What the fuck is there to learn? Before I got sick I had finally gotten my depression under control and had found people who I really loved and was creating a proper life for myself, and nothing is worse than feeling happy for a short time only to have it taken away from you. I had a drinking problem sure (which is the reason I got ill), but I was working to quit drinking already. Those people who I love are the only reason I don't just end it, everytime I've thought about I can only think of how it would affect them and can't go through with it. But right now I can't even comprehend how I'm going to be able to live a life with all this worry and pain, and I don't know what I can do to alleviate it or make it tolerable. Sorry about all the swearing, I'm just in a really bad place right now and know how you feel.

[Check this box to ensure you are paying attention!]

**P2:** I feel ya, I can't purposely kill myself cause I worry about what effects it will have on others.

*I started smoking cigars kind of in hopes that I just like won't wake up one day...*

*I have a liver disorder that leaves scar tissue in my lungs, the doctors always told me if I smoked I'd die. (A little extreme to say something like that to a teen)*

**P1:** I donno a lot of times I think about how it would affect others and feel bad for even considering it in the first place, but the other problem is I've never been open with my friends about my depression and always try to laugh off my illness and other issues in my life. I think they've always known there was a problem because of my heavy drinking and cigarette smoking though, someone even said they feel like they don't know me at all because who I truly am seems so different from what I project to others. I stopped drinking because I was a bitter and abusive drunk though.

And I'm sorry to hear about the liver condition, I'm only 19 (I've lived in the U.K. the last two years so that's how I was able to drink so much) and hearing I have a relatively benign condition that produces painful symptoms was difficult enough to hear, so I can't imagine what it was like to hear about having a life threatening liver condition. I still do smoke even though it makes my symptoms worse, but not nearly as heavily as I used to, because it's the one vice I've always enjoyed. It's just really difficult to hear that your life will never be the same you know? It's not so much anger or sadness but fear of the unknown.

#### Conversation 2

**P1:** It was my understanding that FMLA is something that companies don't really get to 'approve' you for...that if you have the required paperwork and have submitted everything that they 'must' approve you or risk violating certain laws.

One of the basic entitlements for the FMLA is 'to care for a parent who has a serious medical condition.' That should be all you need and usually an HR dept will probably need some paperwork submitted (possibly weekly but it varies) during the leave time to show that you are indeed using that time for caring for your parent. If they give you any grief whatsoever make sure that you mention to your bosses and HR (get all their emails if possible) that you are aware of the federal law and the FMLA says you are entitled to certain rights. They should shape up pretty quickly.

**Edit:** I worked with several people who took FMLA for various reasons and the paperwork were the only hoops they had to jump through. I work for a large corporation as a lowly retail slave. Make sure you keep making noises about your rights and the federal law that protects them. HR gets scared when you know your rights and voice about it. Also make sure you get as much as possible in writing because if they try to screw you they could be in violation and you could have a case for a lawsuit.

**P2:** Agreed but then the question comes up if OP is really 'caring' for his father vs visiting him before he dies. If he is in a hospital he is already being cared for.

**P1:** That shouldn't matter. The law is clear on this: If he needs to be there for his parent then he needs to be there for his parent. What HR will most likely ask for his papers from the father's doctor showing that hospital visits are taking place.

**P2:** 'visiting' isn't the same as 'caring for'. OP must actually be caring for a family member to invoke FMLA protected leave.

<http://rapidlearninginstitute.com/hric/fmla-violation-family>

*I'm not saying he's not, I have no idea, but if he is just visiting then he doesn't qualify.*

**P1:** so I have to fly back to my hometown to take care of him.

So, yeahhh I assume OP is actually taking care of him.

**P2:** The devil is in the details of what that means

[Check this box to ensure you are paying attention!]

**P1:** Yeah I get that and the FMLA paperwork you posted above does outline it pretty specifically so sure. I was just going on what OP posted.

Figure 5: Example Mechanical Turk task - first half where participants were to read two conversations.

Which conversation is better?

Conversation 1  Conversation 2

**Please answer some specific questions about each conversation:**

How engaged are the participants in the conversation?

<p><b>Conversation 1</b></p> <p><input type="radio"/> Neither Engaged</p> <p><input type="radio"/> Only One Engaged</p> <p><input type="radio"/> Slightly Engaged</p> <p><input type="radio"/> Considerably Engaged</p> <p><input type="radio"/> Extremely Engaged</p>	<p><b>Conversation 2</b></p> <p><input type="radio"/> Neither Engaged</p> <p><input type="radio"/> Only One Engaged</p> <p><input type="radio"/> Slightly Engaged</p> <p><input type="radio"/> Considerably Engaged</p> <p><input type="radio"/> Extremely Engaged</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

How supportive is the conversation?

<p><b>Conversation 1</b></p> <p><input type="radio"/> Not At All</p> <p><input type="radio"/> Slightly</p> <p><input type="radio"/> Considerably</p> <p><input type="radio"/> Extremely</p>	<p><b>Conversation 2</b></p> <p><input type="radio"/> Not At All</p> <p><input type="radio"/> Slightly</p> <p><input type="radio"/> Considerably</p> <p><input type="radio"/> Extremely</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

How much do participants seem to be connecting?

<p><b>Conversation 1</b></p> <p><input type="radio"/> Not At All</p> <p><input type="radio"/> Slightly</p> <p><input type="radio"/> Considerably</p> <p><input type="radio"/> Extremely</p>	<p><b>Conversation 2</b></p> <p><input type="radio"/> Not At All</p> <p><input type="radio"/> Slightly</p> <p><input type="radio"/> Considerably</p> <p><input type="radio"/> Extremely</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Was either participant explicitly appreciative of the interaction?

<p><b>Conversation 1</b></p> <p><input type="radio"/> Yes <input type="radio"/> No</p>	<p><b>Conversation 2</b></p> <p><input type="radio"/> Yes <input type="radio"/> No</p>
----------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------

How did they show their appreciation?

Conversation 1 reasons

---

Conversation 2 reasons

---

Any other reasons you found one conversation better than the other?

Optional

---

I've answered all questions! Failure to do so may result in no payment.

[Submit](#)

Figure 6: Example Mechanical Turk task - second half where participants were to answer questions about each conversation.

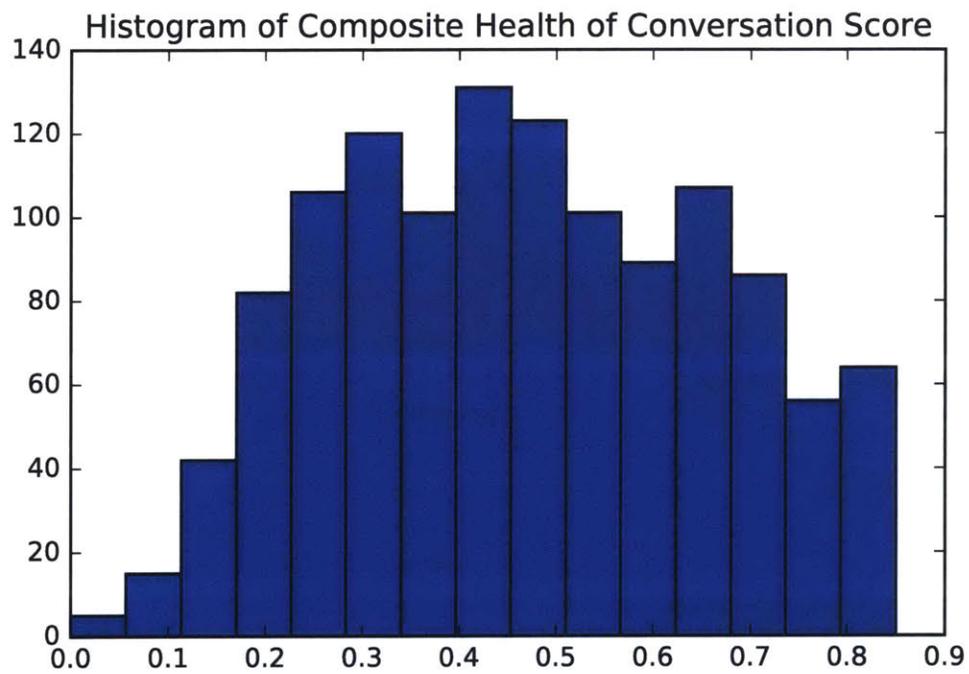


Figure 7: Distribution Health of Conversation Scores for the subset of conversations where all three raters agreed on which conversation was better.

[-] [deleted] 1 point 5 years ago\*

⚡ In neef of a accountability partner ?

permalink embed save

[-] lvshley  over one year 1 point 5 years ago

⚡ countability partner???

permalink embed save parent give award

[-] [deleted] 1 point 5 years ago

⚡ accountability partner sry

permalink embed save parent

---

[-] ginger\_sprout  292 days 3 points 8 months ago

⚡ Hey there, sorry that you're having a hard time. I've quit drinking a couple of times, and the second month was always a little harder for me than the first. There's a thing called [PAWS - Post Acute Withdrawal Syndrome](#). Basically, it takes time for the brain and nervous system to recover from addiction, and you can continue to have periods of feeling like crap for months afterwards. So that's something to be aware of, although I know that I'd be talking to my psychiatrist if I had a sudden uptick in depression and anxiety. So it's good that you've got an appointment coming up.

Basically I'm focusing on self-care and making sure that I'm continuing to support my recovery when I'm feeling less than great. I've been hitting support group meetings and finding it really helps me to get out of my own head and connect with other people. I was isolating pretty hardcore, and it has been nice to meet people who are on the same wavelength with where I'm at. This time around, now that I'm getting more active in my sobriety, I'm meeting some really cool people. I'm an introvert and it does require me going against some of my immediate impulses, but it has been really worthwhile for me to go to support group meetings.

I hope that you find some relief. Be gentle with yourself, this is a big change. IWNDWYT

permalink embed save give award

[-] 7000litres  270 days [S] 2 points 8 months ago

⚡ love your sincere reply  i have dropped aa meeting because gym and work consumes all my day(which is great). Dont really like to be there exsept for youngsters meeting that happens once on fridays ;) . I really need a new weekend hobby which involves a group of people , preferably sober people.

permalink embed save parent give award

[-] ginger\_sprout  292 days 1 point 8 months ago

⚡ Gotcha. Running/walking/hiking groups seem pretty sober. Language/music/art classes are another place where you might meet people. Volunteering. Meetup.com is worth checking out. You might not be able to find one activity that fits the bill, but if gives you a good opportunity to explore some interests. Have fun!

permalink embed save parent give award

Figure 8: Reddit conversations from each end of the spectrum, with health scores of 0.00 (top) and 0.85 (bottom).

# Chapter 5

## Drivers of Healthy Conversation

### 5.1 Feature Engineering

Building off of the research mentioned in the Related Work section, we engineered features from the Reddit conversations to address different potential aspects of conversational health. These broadly fall into three categories:

- Metadata about Conversation
- Homophily of Participants
- Conversational Style

#### 5.1.1 Metadata about Conversation

Four general metadata features were mined:

- Interchanges Count - The total number of conversation turns in the dyad. This ranged from 3 to 43 in the dataset labeled by the mechanical turk workers.
- Average Word Count - The average number of words in the exchange. This is

the number of words per interchange divided by the number of interchanges, and ranged from 1.67 to 629.25.

- Average Words per Sentence - The average number of words per sentence in the whole exchange, which ranged from 1.67 to 47.48.
- Subreddit Category. This is the category of the subreddit that the conversation takes place in, as defined in Table 3.2. The categories are: Mental Health, Relationships, Community, Ending Addiction, and Other.

### 5.1.2 Homophily of Interests

Two homophily features were engineered that related to the similarity of interests of the participants engaging in the conversation:

- Cosine Similarity of Homophily - The similarity of users in terms of the subreddits they had previously participated in (by either posting or commenting). As it is a cosine similarity, the value ranged from 0 to 1.
- Cosine Similarity of Homophily Expanded - The similarity of users in terms of the subreddits they had previously participated in, taking into account similarity of subreddits, as further explained below. As it is a cosine similarity, the value ranged from 0 to 1.

As a motivating example for why mining shared homophily characteristics could contribute positively to promoting healthy conversation between individuals, on one post where someone had shared coping strategies for dealing with loneliness, a conversation had started to emerge between two people. Overall, this conversation went well, with the two trading coping strategies back and forth. However, one individual

mentioned that they like to smoke weed to cope. In response, the other participant became judgmental, scolding them for coping in this manner, and the conversation halted. In this case, the conversation would likely have been more healthy if both individuals shared the same opinions on marijuana.

The first homophily measure was constructed by calculating the number of overlapping subreddits that reddit authors had participated in. However, we were concerned that this method may fall short when comparing individuals who never interacted in the exact same subreddits, but who interacted in similar subreddits, and should therefore have some higher homophily score. For example, a given person may have only interacted with e.g., r/LeagueOfLegends, but we should also be able to associate them with having an interest in the broader category of r/Gaming. One way would be to use an already-defined hierarchy of subreddits<sup>1</sup>. This is a multilevel hierarchy of subreddits. For example, the "Video Game" top level categorization includes below it categories of video game consoles and individual video games, with individual video games being further categorized into the different types of video games. However, if new subreddits are added over time, this would have to be constantly manually updated to stay up to date. Instead, we can algorithmically determine subreddit similarity by taking into account the overlap of redditors between that subreddit and others. We did this using the data collected by Trevor Martin<sup>2</sup>. In short, he analyzed over 1.2 billion comments made by users across 47,494 subreddits (from January 2015 to October 2016) to compute a similarity score between subreddits. We then took the matrix multiplication of the user-subreddit matrix with the subreddit-similarity matrix to compute user vectors of how much they were interested with each subreddit. In this way, expanding the interests of each user. We then calculated the cosine similarity between pairs of user vectors to determine the final homophily score.

---

<sup>1</sup><https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits>

<sup>2</sup><https://www.shorttails.io/interactive-map-of-reddit-and-subreddit-similarity-calculator/>

In summary, the steps we used to calculate a homophily score between participants were:

1. Determine the subreddits the participant had previously interacted with and the number of times they had interacted by posting or commenting. (For the first homophily measure, we calculated the cosine similarity on this matrix).
2. Calculate cosine similarity of all subreddits by taking into account the overlap of users that had interacted with them (using data from Trevor Martin <sup>3</sup>).
3. Expand the user subreddit vector to take into account related subreddits by matrix multiplying the user-subreddit matrix with the subreddit-similarity matrix.
4. Calculate cosine similarity between each user's vector to compute a homophily score for a pair of users.

### 5.1.3 Conversational Style

92 conversational style features were engineered surrounding conversational style from the LIWC [54] library. This includes the following types of features:

- Summary Variables - Analytic, Clout, Authentic, Tone
- Part of Speech - Pronoun, Personal Pronoun, Article, Preposition
- Emotion - Positive Emotion, Negative Emotion, Anxiety, Anger, Sadsness
- Tense - Past, Present, Future
- Punctuation

---

<sup>3</sup><https://www.shorttails.io/interactive-map-of-reddit-and-subreddit-similarity-calculator/>

All of these variables ranged from 0 to 100, and the value was the percent of the words used that mapped to one of LIWC's defined dictionaries of words for each of the categories [54]. LIWC was chosen because, as mentioned in the Related Work section, it has been a central part of psychology research on supportive conversation. We explored using non-LIWC methods to define the summary and sentiment variables, but ultimately decided on LIWC for its popularity amongst psychology research. One method we explored for labeling the Discourse Styles was to use work done by Zhang et al. [70]. They developed a method based on the sociology research in discourse analysis to label conversations on Reddit, and built a labeled model to further classify Reddit conversations. One downside of this was that it was trained on all Reddit data, so did not generalize as well to conversations on a particular topic, and they did not cover the discourse categories as well as LIWC did. The method we explored for sentiment labeling was DeepMoji [19], which, as mentioned in the Related Work section, was trained on Twitter data to predict emotion. Because it was trained on Twitter data, we decided it would be a worse fit for classifying our text.

## 5.2 Results

To analyze which features were most predictive of conversational health, we performed a linear regression with health score as the dependent variable and the 97 features described above as independent variables. Additionally, we performed stepwise model selection by BIC to select the most significant final variables. This method first builds a model with all features, then performs a series of rounds that remove one of the features until removing features no longer achieves a better BIC score. As can be seen in Table 5.1, 13 of the features were selected via model selection as most important to the conversational health score.

Table 5.1: Results of the Linear Regression for predicting Conversational Health

	<i>Dependent variable:</i>
	Health Score (Std. Error)
Word Count	0.001*** (0.0001)
Clout	0.002*** (0.0003)
Authentic	0.001*** (0.0003)
Tone	0.001*** (0.0002)
i	0.011*** (0.002)
prep	0.004** (0.001)
conj	0.008*** (0.002)
posemo	0.004*** (0.001)
time	0.006*** (0.001)
nonflu	-0.014** (0.005)
QMark	-0.006*** (0.002)
interchanges	0.016*** (0.002)
Cosine Similarity - not expanded	0.056*** (0.017)
Constant	-0.035 (0.035)
Observations	1,114
R <sup>2</sup>	0.303
Adjusted R <sup>2</sup>	0.295
Residual Std. Error	0.161 (df = 1101)
F Statistic	39.859*** (df = 12; 1101)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

We performed leave-one-out cross validation to assess the accuracy of our model. To do this, we first selected each pair of conversations that our mechanical turk workers annotated, then built a linear model via the same model selection as described above on the remaining data, then predicted the health score for each conversation. Accuracy was assessed as the percent of time that the better conversation based on the model's predicted scores for each agreed with which conversation the mechanical turk workers agreed was better. We achieved an accuracy of 81.07%.

### 5.3 Discussion

All of the features presented in Table 5.1 are significant at the  $p < .01$  level and are therefore good predictors of conversational health.

It is not surprising that word count and number of interchanges are positively correlated. Given that we saw so few toxic comments, the longer (both in terms of length of writing and number of conversational turns) the conversation, the more likely that people are conversing more, and therefore having a good discussion as both are continuing it. However, it is possible that these results are a result of the mechanical turk task in that conversations that look longer, may be more likely to appear, at face value, better than another conversation. To further evaluate this in the future, we would hold these constant when selecting conversations for mechanical turk workers to compare. In this way they would not be biased by the length.

Interestingly, cosine similarity is significant, but the expanded cosine similarity is not. This indicates that those that have healthier conversations, are also more similar in the subreddits that they talk in. It's possible that taking into account subreddit similarity as a way to gain more information on the reddit user actually diluted the specific interests of a user. It's also possible that users felt some allegiance and connec-

tion to the subreddits they participate in, so if the conversations are happening in those subreddits, they may be more likely to continue conversations, perhaps even “knowing” some of the other posters in the subreddit and conversing more with them. Additionally, it’s likely that having high homophily of interests is not necessarily important when first having a conversation with somebody, but instead, more important for later conversations and maintaining friendships. As a result, general similar interests may be playing a minimal role in how people are communicating.

Emotional words also show up as significant positive influences. Tone (Emotional Tone) and posemo are two features from LIWC that relate to displaying positive emotion. In particular, high Tone relates to a more positive, upbeat writing style while low Tone reveals greater anxiety, sadness, or hostility. posemo looks for words such as “love”, “nice”, or “sweet”. Especially in a depression forum, but also in general, it’s likely that people want to continue the conversation further when supportive words are being included. It also shows a lack of emotionality which may indicate that conversation participants are discussing feelings which is one way to cope.

Clout, Authentic, and personal pronouns are indicators of types of conversations. High clout indicates that the author is speaking from a perspective of high expertise and is confident while a low clout suggests a tentative, humble, or anxious style. Therefore, high clout comments are likely more advice-giving and those doing so with a confident attitude. High authenticity indicates that someone is communicating in an authentic or honest way, oftentimes by being personal and vulnerable, and therefore sharing more personal stories and experiences. Usage of the word “I” and other personal pronouns indicate that someone is sharing a personal experience. This is both sharing personal stories and sharing personal advice. All three are beneficial in healthy supportive conversations.

Another type of conversation is a question-asking one. However, asking a lot of

questions is a negative predictor of conversational health, so asking more questions is actually bad. This variable is a percentage of the total tokens in the interchange that are question marks, so perhaps asking questions is alright, but only if you have additional content around it. Conversations that are just questions are not as supportive of conversations.

Non-fluencies (nonflu) were negatively correlated with conversational health. Non-fluencies that LIWC checks for include hmm, uhh, umm. It's odd that people would write these out on a social media posting as they tend to be more common in real-time speech than a post that is more thought out. It's possible that posts that did include a high number of non-fluencies are ones where the poster wrote his response more hastily, writing as if he were speaking it, and did not put as much thought into the post. Additionally, a poster may have been trying to make clear his uncertainty and lack of experience (somewhat opposite to clout). On the contrary, prepositions and conjunctions were indicators of conversational health. These indicate a level of conversational fluency, especially with conjunctions indicating more complex sentences.

The last significant feature was time. These are words such as "end", "until", "season" which would be used in the context of "this will end soon" or "there will not be much time until you begin feeling better" or other supportive phrases.

# Chapter 6

## Designing an AI System

As seen in the last section, there are many indicators of healthy conversation. In this chapter, we envision an interface that would put this knowledge to use. We aim to create a modified version of Reddit which would more quickly allow posters to find the right person to connect with so that they can have better conversations. This will take into account the variables of homophily and conversational style that were presented in the last chapter. We are mocking this up for Reddit, but the idea is that it could be expanded to other social media or forum sites where people are in need of support (and where there is some amount of user and posting history).

For each Reddit comment, you could annotate it with the homophily and conversational style indicators that we found in the last section. An example is shown in Figure 9. The 5 boxes on the top right are the added annotations. As homophily was a positive indicator of conversational health, we have shown that you have high homophily of interests with the commenter. The two boxes to the right indicate the interests that are shared. In light of the results, perhaps these should be specific subreddits instead of broader interests, but it's possible that with additional data exploration or topic modeling, broader categories of interest could be significant predictors of conversational

health. Additionally, the knowledge of the specific homophily interests may help act as conversation starters. For example, knowing that you both like traveling, you could share a story from a recent trip abroad if it fits in with the supportive conversation. To the right of homophily is an indicator of conversational style of the post. As we found that sharing personal stories, being emotional, and sharing advice (showing expertise or confidence) are positive indicators, this could be highlighted by a positive style and specifying the type of conversation. However, posts that ask too many questions could be flagged as ones with worse conversational style.

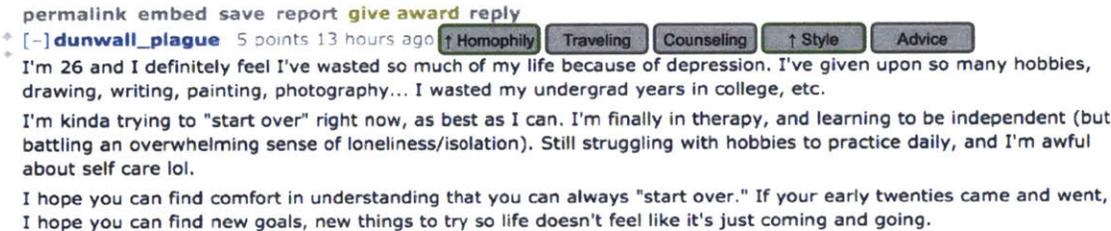


Figure 9: Annotated comments in the interface for indicators of high health of conversation.

Overall, this comment is part of an entire interface of annotated comments, as can be seen in Figure 10. The idea would be that after someone has made a post and a few comments have come in, an interface such as this one could help the one that posted determine who to interact with. Postings could even be sorted via levels of conversational health, with those towards the top predicted to lead to a more healthy conversation.

As is with any system, they often take time to train. The first iteration of the model could be built off of general results of past interactions on Reddit, but in order for the interface to become better and more personalized, a human-in-the-loop design should be implemented. Specifically, research has shown that different attachment styles prefer different types of conversation [47], so the global trends found earlier may

not hold. Age is also a factor in the type of support people prefer to receive, with younger people preferring distractions from their problems while older people prefer rationalization of their choices [15]. To take these into account, those that posted could help train the system as they chose which conversations to interact with. They would be implicitly helping the model improve just via the comments and conversations they select to engage with, but they could additionally explicitly tell the model which of the tagged features they found the most relevant. In this way, the system could learn which measure of homophily are most important to people (and individual commenters) and continue to improve its recommendations.

As with any interface, it is important to keep in mind the information we are bringing forward that, even if public, is now being presented in a new way that may cause concerns. For example, people may be concerned with privacy. Tagging peoples' interests may lead to easier ways to target those that have different beliefs from us. This is something that should be seriously taken in mind while designing as many of the posters in these forums are those that may be extra vulnerable. However, as was seen with the low levels of toxic comments on these postings, these Reddit forums are highly supportive places, so hopefully presenting sensitive information in these contexts would be less likely to be used for harm.

reddit **DEPRESSION** comments

Ever look back and realise you've lost literal years of your life to depression?  
 (self.depression)  
 submitted 17 hours ago by NoxinLimey

Like, I realise this is something that's definitely not a new thing, but in the middle of one of my breakdowns recently I kinda sat back and realised that I've been saying "this has been one of the worst years of my life" for the past 4/5 years? Like, I was first diagnosed when I was 20ish, I'm 23 now, and I don't think I've been truly happy for a period longer than like, 2 months during that time.

It seems like a horrific waste, especially as it hit during my uni years, where every adult around me was telling me how "these will be the best years of your life!!". I've tried to change things up buuuuu I guess this is just who I am now? Idk, I'll probably delete this later but I just needed a second to vent aha. Thanks for listening, I guess?

79 comments share save hide give award report crosspost

all 79 comments  
 sorted by: **best (suggested)** -

[-] **SullenSparrow** 24 points 16 hours ago **Homophily** **GOT** **Funny** **Politics** **Style** **Questioning**  
 Haha yeah I'm 25 and been depressed my whole life. My SO who I'm having issues with just screamed at me "no matter what you can never be happy, you're never happy" its so fucking true. Depression is a slow fucking death if you dont end up offing yourself. Faking happiness is so common, what is the real solution to this problem? Can those of us that suffer find true happiness? Most say "yes" but do they know what it's like to wish you were dead every single day? Sorry you're in the same boat hope you find the answer and find some peace soon.  
 permalink embed save report give award reply

[-] **dreamgloss** 12 points 16 hours ago **Homophily** **Female** **Depression** **Style** **Musings**  
 I think of this now and then.  
 I realize that it was out of my control. While I would take the years back in a heartbeat...I wouldn't change a thing. Not because I want to be depressed...or whatever else is wrong with me.  
 The experiences made me, me.  
 Now I'm learning to move past that.  
 Now I'm in treatment.  
 I'm looking forward to the day I can think back to myself and truly believe that I won the battle.  
 The battle of becoming myself. The realization of who that is. Then...just being.  
 permalink embed save report give award reply

[-] **northernsky-** 10 points 15 hours ago **Homophily** **Depression** **Style** **Storytelling**  
 I'm in the same room where I spent my childhood and I'm nearing 30.  
 I've wasted 8 years to suicidal ideation. These years are gone and if I could turn back time, I still wouldn't know how I could have been doing something 'worthwhile', because I never felt I have a calling in life and I couldn't force myself to do something completely random. I've started school - university and trade school 3 times and longest I lasted was probably few months. Maybe it was my fate to be lost for nearly 10 years, I sometimes think that maybe I'm the person who wakes up suddenly to some different perception, but I haven't so far and seems like it's never going to happen. I've never felt good in my own skin, how can I DO something?  
 Lately I obsess over thought that my heart is so closed to life (so I was reading about heart chakra haha understand how desperate people fall into spiritual path), no energy flows through it, through my body, through my heart, I mean what energy can you feel when you don't give anyone anything because you're talentless, when you don't find meaning in things that people do, when you absolutely don't understand why the fuck you were born. Yeah years feel like wasted, we were busy with our anxieties, fears, suicidal ideation, lack of purpose, all different kinds of hell that is possible to experience... and I wish I could end this rant with a perfect solution and conclusion, but the thing is, I'm clueless why it happened and continue to happen with me. I have no wisdom about life. Emptiness leads me towards my death, because there is nothing else to lead me.  
 permalink embed save report give award reply

[-] **dunwall\_plague** 5 points 13 hours ago **Homophily** **Traveling** **Counseling** **Style** **Advice**  
 I'm 26 and I definitely feel I've wasted so much of my life because of depression. I've given upon so many hobbies, drawing, writing, painting, photography... I wasted my undergrad years in college, etc.  
 I'm kinda trying to "start over" right now, as best as I can. I'm finally in therapy, and learning to be independent (but battling an overwhelming sense of loneliness/isolation). Still struggling with hobbies to practice daily, and I'm awful about self care lol.  
 I hope you can find comfort in understanding that you can always "start over." If your early twenties came and went, I hope you can find new goals, new things to try so life doesn't feel like it's just coming and going.  
 permalink embed save report give award reply

[-] **dwemerknicht** 6 points 17 hours ago **Homophily** **Depression** **Memes** **Style** **Storytelling**  
 I will be 29 in August and yup it doesn't get better lol

Figure 10: Modified Reddit interface with annotations for conversational health.

# Chapter 7

## Conclusion

In conclusion, through this work, we were able to redefine a score for health of non-toxic, online conversations that was built on research in psychology. This was a composite measure that took into account supportiveness, engagement, appreciation, and connection between two people conversing online about loneliness. Using that measure, we were then able to begin to understand some of the components that drive conversational health. We found that those engaging that have higher homophily (in terms of the number of subreddits they have in common) are more likely to engage in a healthier conversation. Additionally, we found that certain conversational styles are better for higher conversational health. These types of conversations include sharing personal stories, being emotional, and sharing advice. However, conversations that asked too high a proportion of questions were less correlated with healthy conversation. We finished by envisioning what an interface would look like that took these drivers of healthy conversation into account.

For future work, we would like to prioritize further examining homophily measures. Another way to compute homophily is through analyzing the content of what the Reddit posters have discussed in the past. This could be done through a topic modeling

of the users posts (such as through latent dirichlet allocation [4]). Another method could involve creating a user to vector method [41, 69, 71] taking into account topics discussed and subreddits posted in, then computing the cosine similarity between users as a measure of homophily. Additionally, we would like to investigate homophily of speaking style. This research focused on what types of styles are generally best, but perhaps there are also homophily tendencies in the way people prefer to converse.

The end goal of this research is to help those that are feeling lonely quickly find the best conversation partner who will be able to help them through their situation while having productive conversations. We hope that this thesis research is a step in that direction, so the many millions around the world who are feeling lonely, can soon begin to feel less so.

# Bibliography

- [1] Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 13–22. ACM, 2013.
- [2] Mark W Becker, Reem Alzahabi, and Christopher J Hopwood. Media multitasking is associated with symptoms of depression and social anxiety. *Cyberpsychology, Behavior, and Social Networking*, 16(2):132–135, 2013.
- [3] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. A study of homophily on social media. *World Wide Web*, 15(2):213–232, 2012.
- [4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [5] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- [6] Matthew E Brashears. Small networks and high isolation? a reexamination of american discussion networks. *Social Networks*, 33(4):331–341, 2011.
- [7] Kaitlin Cannava and Graham D Bodie. Language use and style matching in supportive conversations between strangers and friends. *Journal of Social and Personal Relationships*, 34(4):467–485, 2017.
- [8] Scott E Caplan. Relations among loneliness, social anxiety, and problematic internet use. *CyberPsychology & behavior*, 10(2):234–242, 2006.
- [9] Shuo Chang, Vikas Kumar, Eric Gilbert, and Loren G Terveen. Specialization, homophily, and gender in a social curation site: findings from pinterest. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 674–686. ACM, 2014.
- [10] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

- [11] Robert Crosnoe. Friendships in childhood and adolescence: The life course and new directions. *Social psychology quarterly*, pages 377–391, 2000.
- [12] Sergio Currarini, Matthew O Jackson, and Paolo Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.
- [13] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754. ACM, 2011.
- [14] Felix R Day, Ken K Ong, and John RB Perry. Elucidating the genetic basis of social interaction and isolation. *Nature communications*, 9(1):2457, 2018.
- [15] Kathy Denton and Lynne Zarbatany. Age differences in support processes in conversations between friends. *Child Development*, 67(4):1360–1373, 1996.
- [16] Nicholas Diakopoulos and Mor Naaman. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 133–142. ACM, 2011.
- [17] Gabriel Doyle, Dan Yurovsky, and Michael C Frank. A robust framework for estimating linguistic alignment in twitter conversations. In *Proceedings of the 25th international conference on world wide web*, pages 637–648. International World Wide Web Conferences Steering Committee, 2016.
- [18] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.
- [19] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.
- [20] Nicholas FitzGerald, Giuseppe Carenini, Gabriel Murray, and Shafiq Joty. Exploiting conversational features to detect high-quality blog comments. In *Canadian Conference on Artificial Intelligence*, pages 122–127. Springer, 2011.
- [21] Devin Gaffney and J Nathan Matias. Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *arXiv preprint arXiv:1803.05046*, 2018.
- [22] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):3, 2018.
- [23] Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. Convolutional neural networks for twitter text toxicity analysis. In *INNS Big Data and Deep Learning conference*, pages 370–379. Springer, 2019.

- [24] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.
- [25] Geir Godager. Birds of a feather flock together: A study of doctor–patient matching. *Journal of health economics*, 31(1):296–305, 2012.
- [26] Jill V Hamm. Do birds of a feather flock together? the variable bases for african american, asian american, and european american adolescents’ selection of similar friends. *Developmental psychology*, 36(2):209, 2000.
- [27] Keith N Hampton, Lee Rainie, Weixu Lu, Inyoung Shin, and Kristen Purcell. Social media and the cost of caring. *Washington, DC: Pew Research Center*, 2015.
- [28] Steven J Heine, Julie-Ann B Foster, and Roy Spina. Do birds of a feather universally flock together? cultural variation in the similarity-attraction effect. *Asian Journal of Social Psychology*, 12(4):247–258, 2009.
- [29] Susan C Herring. The coevolution of computer-mediated communication and computer-mediated discourse analysis. In *Analyzing Digital Discourse*, pages 25–67. Springer, 2019.
- [30] Melissa G Hunt, Rachel Marx, Courtney Lipson, and Jordyn Young. No more fomo: Limiting social media decreases loneliness and depression. *Journal of Social and Clinical Psychology*, 37(10):751–768, 2018.
- [31] Susanne M Jones and Graham D Bodie. 16 supportive communication. *Interpersonal communication*, 6:371, 2014.
- [32] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*, 2017.
- [33] Lauri Kovanen, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proceedings of the National Academy of Sciences*, page 201307941, 2013.
- [34] Brian Lakey, Jana Brittain Drew, and Kimberly Sirl. Clinical depression and perceptions of supportive others: A generalizability analysis. *Cognitive Therapy and Research*, 23(5):511–533, 1999.
- [35] Elliot A Layden, John T Cacioppo, Stephanie Cacioppo, Stefano F Cappa, Alessandra Dodich, Andrea Falini, and Nicola Canessa. Perceived social isolation is associated with altered functional connectivity in neural networks associated with tonic alertness and executive control. *Neuroimage*, 145:58–73, 2017.
- [36] Campbell Leaper, Mary Carson, Carilyn Baker, Heithre Holliday, and Sharon Myers. Self-disclosure and listener verbal support in same-gender and cross-gender friends’ conversations. *Sex Roles*, 33(5-6):387–404, 1995.

- [37] Alex Leavitt and Joshua A Clark. Upvoting hurricane sandy: event-based news production processes on a social news site. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1495–1504. ACM, 2014.
- [38] Alex Leavitt and John J Robinson. The role of information visibility in network gatekeeping: Information aggregation on reddit during crisis events. In *CSCW*, pages 1246–1261, 2017.
- [39] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.
- [40] Liu Yi Lin, Jaime E Sidani, Ariel Shensa, Ana Radovic, Elizabeth Miller, Jason B Colditz, Beth L Hoffman, Leila M Giles, and Brian A Primack. Association between social media use and depression among us young adults. *Depression and anxiety*, 33(4):323–331, 2016.
- [41] Haiying Liu, Lifang Wu, Dai Zhang, Meng Jian, and Xiuzhen Zhang. Multi-perspective user2vec: Exploiting re-pin activity for user representation learning in content curation social network. *Signal Processing*, 142:450–456, 2018.
- [42] Adrienne Massanari. *Participatory Culture, Community, and Play. Chapter 2. Defining reddit*. Peter Lang, Bern, Switzerland.
- [43] Adrienne Massanari. #gamergate and the fapping: How reddit’s algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3):329–346, 2017.
- [44] Julian McAuley and Alex Yang. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee, 2016.
- [45] Miller McPherson, Lynn Smith-Lovin, and Matthew E Brashears. Social isolation in america: Changes in core discussion networks over two decades. *American sociological review*, 71(3):353–375, 2006.
- [46] Todor Mihaylov and Preslav Nakov. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 399–405, 2016.
- [47] Mario Mikulincer and Victor Florian. Are emotional and instrumental supportive interactions beneficial in times of stress? the impact of attachment style. *Anxiety, stress, and coping*, 10(2):109–127, 1997.

- [48] Courtney Napoles, Aasish Pappu, and Joel Tetreault. Automatically identifying good conversations online (yes, they do exist!). In *Eleventh International AAAI Conference on Web and Social Media*, 2017.
- [49] Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23, 2017.
- [50] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.
- [51] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.
- [52] Albert Park and Mike Conway. Longitudinal changes in psychological states in online health community members: understanding the long-term effects of participating in an online depression community. *Journal of medical Internet research*, 19(3), 2017.
- [53] Albert Park and Mike Conway. Harnessing reddit to understand the written-communication challenges experienced by individuals with mental health disorders: Analysis of texts from mental health communities. *Journal of medical Internet research*, 20(4), 2018.
- [54] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.
- [55] Robert Plutchik. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219, 1984.
- [56] Yuqing Ren, F Maxwell Harper, Sara Drenner, Loren Terveen, Sara Kiesler, John Riedl, and Robert E Kraut. Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *Mis Quarterly*, pages 841–864, 2012.
- [57] Jane R Rosen-Grandon, Jane E Myers, and John A Hattie. The relationship between marital characteristics, marital interaction processes, and marital satisfaction. *Journal of Counseling & Development*, 82(1):58–68, 2004.

- [58] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, 2017.
- [59] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*, 2017.
- [60] Wesley Shrum, Neil H Cheek Jr, and Saundra MacD. Friendship in school: Gender and racial homophily. *Sociology of Education*, pages 227–239, 1988.
- [61] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1631, page 1642, 2013.
- [62] Mike Thelwall. Homophily in myspace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231, 2009.
- [63] Teun A Van Dijk. Discourse analysis: Its development and application to the structure of news. *Journal of communication*, 33(2):20–43, 1983.
- [64] Janne Vanhalst, Brandon E Gibb, and Mitchell J Prinstein. Lonely adolescents exhibit heightened sensitivity for facial cues of emotion. *Cognition and emotion*, 31(2):377–383, 2017.
- [65] Duncan J Watts, Peter Sheridan Dodds, and Mark EJ Newman. Identity and search in social networks. *science*, 296(5571):1302–1305, 2002.
- [66] Robert L Weiss and Kendra J Summers. Marital interaction coding system-iii. *Marriage and family assessment*, pages 85–115, 1983.
- [67] Heather Cleland Woods and Holly Scott. # sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *Journal of adolescence*, 51:41–49, 2016.
- [68] K Lira Yoon and Richard E Zinbarg. Interpreting neutral faces as threatening is a default mode for socially anxious individuals. *Journal of abnormal psychology*, 117(3):680, 2008.
- [69] Yang Yu, Xiaojun Wan, and Xinjie Zhou. User embedding for scholarly microblog recommendation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 449–453, 2016.
- [70] Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. Characterizing online discussion using coarse discourse sequences. In *Proceedings of the Eleventh International Conference on Web and Social Media*. AAAI Press, 2017.

- [71] Konrad Zolna and Bartłomiej Romański. User modeling using lstm networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017.