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

Signature redacted

Program in Media Arts and Sciences
May 24, 2019

Certified by

Associate Professor, Program in Media Arts and Sciences

Accepted by

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 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, Pranjal, 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

1 Introduction 10

2 Related Work 15
   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

3 Data 20

4 Defining Healthy Conversation 25
   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

5 Drivers of Healthy Conversation .......................... 39
  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

6 Designing an AI System ................................... 48

7 Conclusion .................................................... 52
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\(^1\). 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].

\(^1\)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/psuedo-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
I hate days like Halloween

They remind me of how so many other people get to go out and have fun alone by myself with no friends to see just another day to stay at home an and cry. Everyone else has someone else that they like more than me, I'm alone that's why I have nowhere to go. Probably gonna be the same next year.

How do you accept that you're actually depressed?

At what point do you realize that you were depressed and need to accept it?

How do you validate your thoughts and feelings?

What made you finally click and say: "I need to get help?"

I hate days like Halloween.

I'm 21 and not in school. I used to enjoy video games, but since I was a little kid. I have no friends. Only a long distance girlfriend who doesn't share the same interest as me and I feel like I'm pushing her away.

I spent all my free time playing games and watching anime when I'm not at work. I can't even enjoy this stuff anymore. I feel like I need to explain it like I should be doing more with my life, making my parents proud that playing games and looking at anime women. I don't know what I want to do on the TV, it's so empty that it's just silent. The answer for me never seems to come from anywhere. It's empty and empty and empty.

I scheduled my first appointments to get better.

Hi everyone.

Today I had the courage to walk into my university's health appointments. One for therapy and the other for medical evaluation and have been having temporary suicidal thoughts for the past few months. This is a hard decision, but I know it's the right one.

I've been feeling lost and unmotivated for too long. It's time for a change.

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.
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\(^1\) 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

\(^1\)reddit.com
other fields that are available through Reddit’s API\(^2\). Since then, the entire dataset has been uploaded to Google’s BigQuery\(^3\) 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.

\(^2\)https://www.reddit.com/r/datasets/comments/3bx1g7/i_have_every_publicly_available_reddit_comment/
Table 3.1: Total posts and comments on the top 20 represented subreddits.

<table>
<thead>
<tr>
<th>SUBREDDIT</th>
<th>TOTAL POSTS</th>
<th>TOTAL COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>depression</td>
<td>632,697</td>
<td>2,783,807</td>
</tr>
<tr>
<td>relationships</td>
<td>84,916</td>
<td>1,259,167</td>
</tr>
<tr>
<td>offmychest</td>
<td>69,189</td>
<td>262,310</td>
</tr>
<tr>
<td>askreddit</td>
<td>65,218</td>
<td>1,595,738</td>
</tr>
<tr>
<td>suicidewatch</td>
<td>56,250</td>
<td>351,913</td>
</tr>
<tr>
<td>nofap</td>
<td>49,174</td>
<td>362,914</td>
</tr>
<tr>
<td>relationship_advice</td>
<td>35,449</td>
<td>340,811</td>
</tr>
<tr>
<td>advice</td>
<td>28,746</td>
<td>142,560</td>
</tr>
<tr>
<td>r4r</td>
<td>27,312</td>
<td>76,721</td>
</tr>
<tr>
<td>anxiety</td>
<td>24,750</td>
<td>122,397</td>
</tr>
<tr>
<td>lonely</td>
<td>22,588</td>
<td>123,109</td>
</tr>
<tr>
<td>raisedbynarcissists</td>
<td>22,147</td>
<td>217,887</td>
</tr>
<tr>
<td>bipolar</td>
<td>21,453</td>
<td>163,445</td>
</tr>
<tr>
<td>dirtypenpals</td>
<td>20,201</td>
<td>9,247</td>
</tr>
<tr>
<td>stopdrinking</td>
<td>16,611</td>
<td>199,060</td>
</tr>
<tr>
<td>mentalhealth</td>
<td>15,261</td>
<td>58,392</td>
</tr>
<tr>
<td>drugs</td>
<td>14,787</td>
<td>263,540</td>
</tr>
<tr>
<td>trees</td>
<td>14,397</td>
<td>160,828</td>
</tr>
<tr>
<td>teenagers</td>
<td>13,912</td>
<td>178,227</td>
</tr>
<tr>
<td>needafriend</td>
<td>13,661</td>
<td>55,669</td>
</tr>
</tbody>
</table>

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 lonliness 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.

<table>
<thead>
<tr>
<th>Subreddit</th>
<th>Mental Health</th>
<th>Relationships</th>
<th>Community</th>
<th>Ending Addiction</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>depression</td>
<td>relationship_advice</td>
<td>relationship_advice</td>
<td>asktransgender</td>
<td>NoFap</td>
<td>offmychest</td>
</tr>
<tr>
<td>SuicideWatch</td>
<td>raisedbynarcissists</td>
<td></td>
<td></td>
<td>stopdrinking</td>
<td>advice</td>
</tr>
<tr>
<td>Anxiety</td>
<td>foreveralone</td>
<td></td>
<td></td>
<td>loseit</td>
<td>drugs</td>
</tr>
<tr>
<td>Lonely</td>
<td>breakups</td>
<td>deadbedrooms</td>
<td>askgaybros</td>
<td>leaves</td>
<td>casualconversation</td>
</tr>
<tr>
<td>Bipolar</td>
<td>dating_advice</td>
<td>unsentletters</td>
<td>askmen</td>
<td>pornfree</td>
<td>mmfb</td>
</tr>
<tr>
<td>Mentalhealth</td>
<td>longdistance</td>
<td>exnochontact</td>
<td>keto</td>
<td>nootropics</td>
<td>needadvice</td>
</tr>
<tr>
<td>Depression_help</td>
<td></td>
<td>Divorce</td>
<td>fitness</td>
<td>stopsmoking</td>
<td>trueoffmychest</td>
</tr>
<tr>
<td>ADHD</td>
<td></td>
<td></td>
<td>breakingmom</td>
<td></td>
<td>rant</td>
</tr>
<tr>
<td>BPD</td>
<td></td>
<td></td>
<td>actuallesbians</td>
<td></td>
<td>personalfinance</td>
</tr>
<tr>
<td>BipolarReddit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>asktrp</td>
</tr>
<tr>
<td>Socialanxiety</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>jobs</td>
</tr>
<tr>
<td>Depressed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspergers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selfharm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socialskills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Getting_over_it</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depressionregimens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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\(^1\).

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

\(^1\)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. 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-
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

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

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.

<table>
<thead>
<tr>
<th>PERSPECTIVE API SCORE</th>
<th>TOXICITY</th>
<th>OBSCENE</th>
<th>INFLAMMATORY</th>
<th>COMMENT</th>
<th>COMMENT ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0057</td>
<td>0.0062</td>
<td>0.0146</td>
<td></td>
<td>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.</td>
<td>DJVEAAW</td>
</tr>
<tr>
<td>0.0479</td>
<td>0.3372</td>
<td>0.3679</td>
<td></td>
<td>DUNNO IF I CAN START A GAMING CLUB THIS EGYPTIAN UNI :P</td>
<td>C69LQQW</td>
</tr>
<tr>
<td>0.2515</td>
<td>0.9661</td>
<td>0.6995</td>
<td></td>
<td>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.</td>
<td>CLJ15K0</td>
</tr>
<tr>
<td>0.3134</td>
<td>0.0198</td>
<td>0.2662</td>
<td></td>
<td>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?</td>
<td>DWL207J</td>
</tr>
<tr>
<td>0.3601</td>
<td>0.6574</td>
<td>0.7079</td>
<td></td>
<td>SOUNDS LIKE YOU IDENTIFIED YOUR PROBLEM. THAT'S A GOOD FIRST STEP.</td>
<td>DZ419QP</td>
</tr>
<tr>
<td>0.4985</td>
<td>0.0227</td>
<td>0.7079</td>
<td></td>
<td>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 PROBLEM, AND IF NOT, WELL NOW YOU'VE GOT A GOAL ASSOCIATED WITH YOUR MISERABLE JOB. GOOD LUCK!</td>
<td>C95BLMN</td>
</tr>
<tr>
<td>0.6701</td>
<td>0.6574</td>
<td>0.7236</td>
<td></td>
<td>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 MORNING. SAME AS EVERYONE!</td>
<td>CLU8EGX</td>
</tr>
<tr>
<td>0.7431</td>
<td>0.9833</td>
<td>0.6872</td>
<td></td>
<td>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.</td>
<td>C8RL328</td>
</tr>
<tr>
<td>0.8341</td>
<td>0.9937</td>
<td>0.5967</td>
<td></td>
<td>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 HOP SHE FLIPS HER SHIT. SO I CALL THE COP AND HAVE HER ASS TAKEN AWAY. MAYBE PUT MY INVESTMENTS AND PAY FOR A THERAPIST TO EVALUATE HER AND USE IT IN COURT.</td>
<td>UPNWTP</td>
</tr>
<tr>
<td>0.9257</td>
<td>0.9937</td>
<td>0.4205</td>
<td></td>
<td>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</td>
<td>E6WX2K3</td>
</tr>
<tr>
<td>0.9907</td>
<td>0.9937</td>
<td>0.3599</td>
<td></td>
<td>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! F<em>CKING STUPID! I AM SOOOOO SORRY FOR YOUR SITUATION HOWEVER YOU SEEM LIKE A TOUCH DOZE SO I KNOW YOU'LL BE SMART AND NOT TAKE THAT F</em>CKING SHIT FROM NOBODY! UP VOTES FOR YOU!!!!</td>
<td>CEUJEG3</td>
</tr>
<tr>
<td>0.9930</td>
<td>0.9833</td>
<td>0.4461</td>
<td></td>
<td>YOU'RE A F*CKING MORON.</td>
<td>CBJNODB</td>
</tr>
</tbody>
</table>
Table 4.3: A Sample of the most toxic Reddit Comments with their respective Perspective Toxicity Scores

<table>
<thead>
<tr>
<th>PERSPECTIVE API SCORE</th>
<th>TOXICITY</th>
<th>OBSCENE</th>
<th>INFLAMMATORY</th>
<th>COMMENT</th>
<th>COMMENT ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9979</td>
<td>0.9927</td>
<td></td>
<td>0.4459</td>
<td>FUCK YOU YOU STUPID FAT UGLY GAY SACK OF SHIT</td>
<td>CNKVXTD</td>
</tr>
<tr>
<td>0.9964</td>
<td>0.9920</td>
<td></td>
<td>0.3149</td>
<td>You’re a fucking asshole OP. Fuck you.</td>
<td>C71AMJG</td>
</tr>
<tr>
<td>0.9912</td>
<td>0.9920</td>
<td></td>
<td>0.2675</td>
<td>Go fuck yourself.</td>
<td>CWTH14J</td>
</tr>
<tr>
<td>0.9907</td>
<td>0.9937</td>
<td></td>
<td>0.3882</td>
<td>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 hear you’re tough as well. I wish nothing but the best for you!</td>
<td>C34VKFG</td>
</tr>
<tr>
<td>0.9898</td>
<td>0.9923</td>
<td></td>
<td>0.4501</td>
<td>Shut the fuck up unfunny idiot</td>
<td>DVMSJ3L</td>
</tr>
<tr>
<td>0.9892</td>
<td>0.9892</td>
<td></td>
<td>0.3639</td>
<td>Fuck off</td>
<td>DB00Y79</td>
</tr>
<tr>
<td>0.9839</td>
<td>0.9920</td>
<td></td>
<td>0.3667</td>
<td>Fuck that guy.</td>
<td>E4H6DBO</td>
</tr>
<tr>
<td>0.9838</td>
<td>0.9937</td>
<td></td>
<td>0.4700</td>
<td>0 to you. Fuck those assholes.</td>
<td>DYTCK5I</td>
</tr>
<tr>
<td>0.9838</td>
<td>0.9937</td>
<td></td>
<td>0.4385</td>
<td>Someone downvoted you, fucking assholes. Nah, I appreciate you. Truly.</td>
<td>DSYD0JG</td>
</tr>
<tr>
<td>0.9837</td>
<td>0.9892</td>
<td></td>
<td>0.5527</td>
<td>Fuck that bitch, keep doing you and continue kicking ass and get what you are pursuing.</td>
<td>D6N6K00</td>
</tr>
<tr>
<td>0.9729</td>
<td>0.9918</td>
<td></td>
<td>0.3664</td>
<td>Fuck that guy, you deserve better</td>
<td>COKVGN3</td>
</tr>
<tr>
<td>0.9687</td>
<td>0.9910</td>
<td></td>
<td>0.4537</td>
<td>That is a fucking awful workplace</td>
<td>D62AH5Q</td>
</tr>
</tbody>
</table>
Table 4.4: Percent of Perspective API Scores about score thresholds.

<table>
<thead>
<tr>
<th>SCORE</th>
<th>TOXICITY</th>
<th>OBSCENE</th>
<th>INFLAMMATORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>.95</td>
<td>1.7286%</td>
<td>13.2724%</td>
<td>0.1046%</td>
</tr>
<tr>
<td>.90</td>
<td>0.5635%</td>
<td>14.5276%</td>
<td>0.0006%</td>
</tr>
</tbody>
</table>

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., \( P_1 > P_2 > P_1 \)). 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 question 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

<table>
<thead>
<tr>
<th>AGREEMENT PERCENTAGE</th>
<th>All Scores</th>
<th>w/o Engagement</th>
<th>w/o Support</th>
<th>w/o Connection</th>
<th>w/o Appreciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>All 3 agreed</td>
<td>93.76%</td>
<td>93.82%</td>
<td>93.82%</td>
<td>93.82%</td>
<td>93.70%</td>
</tr>
<tr>
<td>General</td>
<td>83.04%</td>
<td>83.29%</td>
<td>83.77%</td>
<td>83.09%</td>
<td>83.92%</td>
</tr>
</tbody>
</table>

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.

Thank you really appreciate you helping me research on conversational health! - Lauren

This HIT is part of a MFT 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 the HIT voluntarily. Please email Lauren: Fatamico at fatamico@umd.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 feeling that you have. It's really hard, 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 first got depressed (it really ruined my reputation and when I'm visiting home people 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 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 keeps 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 it), but I was working to quit drinking already. Those people who I love are the only reason I don't just end it, everyday 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 don't worry 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 lie 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 smoke I'd die. (A little extreme to say something like that to a teen)

P1: I do now 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 illnesses 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.

P2: so I have to fly back to my hometown to take care of him.

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.

P3: '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/hr/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, yeahh 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.

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 paper work 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 try to screw you they could be in violation and you could have a case for a lawsuit.

P2: Agree'd but then the question comes up if CP 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'. CP must actually be caring for a family member to invoke FMLA protected leave.

http://rapidlearninginstitute.com/hr/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, yeahh I assume CP 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: I was actually talking about the 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.
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.
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.

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 subreddits¹. This is a multilevel hierarchy of subreddits. For example, the "Video Game" top level categorization includes below it categories of video came 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 Martin². 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.

¹https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits
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 \(^3\)).

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, Sadness
- **Tense** - Past, Present, Future
- **Punctuation**

\(^3\)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

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Health Score (Std. Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Count</td>
<td>0.001*** (0.0001)</td>
</tr>
<tr>
<td>Clout</td>
<td>0.002*** (0.0003)</td>
</tr>
<tr>
<td>Authentic</td>
<td>0.001*** (0.0003)</td>
</tr>
<tr>
<td>Tone</td>
<td>0.001*** (0.0002)</td>
</tr>
<tr>
<td>i</td>
<td>0.011*** (0.002)</td>
</tr>
<tr>
<td>prep</td>
<td>0.004** (0.001)</td>
</tr>
<tr>
<td>conj</td>
<td>0.008*** (0.002)</td>
</tr>
<tr>
<td>posemo</td>
<td>0.004*** (0.001)</td>
</tr>
<tr>
<td>time</td>
<td>0.006*** (0.001)</td>
</tr>
<tr>
<td>nonflu</td>
<td>-0.014** (0.005)</td>
</tr>
<tr>
<td>QMark</td>
<td>-0.006*** (0.002)</td>
</tr>
<tr>
<td>interchanges</td>
<td>0.016*** (0.002)</td>
</tr>
<tr>
<td>Cosine Similarity - not expanded</td>
<td>0.056*** (0.017)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.035 (0.035)</td>
</tr>
</tbody>
</table>

Observations: 1,114  
R²: 0.303  
Adjusted R²: 0.295  
Residual Std. Error: 0.161 (df = 1101)  
F Statistic: 39.859*** (df = 12; 1101)

Note: *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” of “there will not me 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.
I'm kinda trying to "start over" right now, as best as I can. 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 buuuut I guess this is just who I am now? Idk, I'll probably delete this later but I just needed a second to vent ahah. Thanks for listening, I guess.

I've wasted so much of my life because of depression. I've given up so many hobbies, drawing, writing, painting, photography... I wasted my undergrad years in college, etc. 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, leads me towards Emptiness... 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.

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


[31] Susanne M Jones and Graham D Bodie. 16 supportive communication. *Interpersonal communication, 6:371, 2014.*


