Using machine-learning to measure empathy in writing

Zhilin Wang
University of Cambridge

For the partial fulfillment of the BA (Hons.)

Supervised by:
Pablo Torres Núñez

Abstract

The development of empathy is a critical aspect of every individual's social emotional learning. Empathy needs to be measured accurately for individuals and those related to them to use this information to determine how empathy can be best developed. Machine-learning based analysis of writing is a highly promising approach of measuring empathy. However, the difficulty of labeling empathetic writing is a key bottleneck that prevents machine-learning-based analysis from fulfilling its promise. This study proposes and validates a novel method of labeling whether writing is empathetic. This method uses the effective helpfulness of people's written help to others, as rated by the recipient of the written help, to label whether writing is empathetic.. This research and investigation project found this labeling method to have good validity in multiple dimensions. Furthermore, the study applies this labeling method in machine-learning analysis of empathetic writing. The exploration reveals that empathetic writing includes language that is positive and friendly, attempting to understand the perspective of others and indicating an absence of elaborate argumentation Overall, this study contributes a novel method that improves machine-learning-based analysis of empathy in writing. In the future, this new measurement method could facilitate an automatic, low-cost and ubiquitous assessment of empathy and, as a consequence, promote effective social emotional learning for all.

Acknowledgements

This project would not have been possible without the generous and patient support of my supervisor, family and friends. Thank you!

Table of Contents

List of Figures

List of Tables

List of Equations

1 Introduction

Empathy is defined as understanding a person from their perspective rather than one's own and indirectly experiencing a person's feelings, perceptions, and thoughts (American Psychological Association [APA], 2020). Such a definition acknowledges the importance of both affective and cognitive aspects of empathy, as stressed by Davis (1983), Zhou, Valiente and Eisenberg (2003) and Baron-Cohen and Wheelwright (2004). The affective component of empathy concerns appreciating the emotional state others are in and provide adequate emotional responses to others (Hoffman, 2000) while the cognitive components centers around understanding how others are likely to think based on their experiences rather than one's own (Leslie, 1987).

Empathy is an important social emotional skill for individuals and society. Not only does empathy motivate pro-social behavior (Eisenberg, Fabes & Spinrad, 2007; Jensen, 2016), it also provides a sense of connection (Zhou et al., 2003). Furthermore, empathy is an aspect of emotional intelligence (Goleman, 1998) as well as interpersonal intelligence (Gardner, 1993), which facilitates social competence. Therefore, the development of empathy is a significant aspect of social emotional learning for everyone.

Given the importance of developing empathy, measuring empathy has become critical. Measuring empathy can provide information about an individual's level of empathy through operationalizing theoretical understandings of empathy. With such information, a person can better understand how they can develop their level of empathy through purposeful educational interventions (Császár, 2012; Kelm, Womer, Walter & Feudtner, 2014; Lam, Kolomitro & Alamparambil, 2011; Srivastava & Das, 2016; Teding van Berkhout & Malouff, 2016). This information can also be provided to family, friends, teachers or counselors given that an individual's level of empathy might not be apparent to them (Batt-Rawden, Chisolm, Anton & Tabor, 2013; Casale, Thomas & Simmons, 2018; Verhofstadt et al., 2016). Knowing a person's level of empathy can enable others close to the person to interact more effectively with them and support their development of empathy (Howe, 2013; Upshaw, Kaiser & Sommerville, 2015).

Furthermore, measuring empathy as expressed in written communication can also support students to provide more empathetic peer feedback to one another in educational processes (Liu & Carless, 2006). Teachers can also make use of such measurements to learn to provide more empathetic and emotionally supportive feedback to students (Warren, 2018). This is especially relevant in a time when online learning becomes more popular, both generally (Palvia et al., 2018) and particularly in this COVID-19 period when schools around the world are forced to shut (Viner et al., 2020).

However, existing methods of measuring empathy have methodological limitations. The literature review section will first evaluate conventional methods used to measure empathy before discussing current progress in measuring empathy through machine-learning analysis of writing. To overcome these methodological limitations, this study proposes and tests a novel approach to measure empathy based on machine-learning analysis of people's writing.

2 Literature Review

2.1 Conventional methods of measuring empathy

Empathy is commonly measured through empathy-invoking situations, performance tasks and self-reported questionnaires.

2.1.1 Analysis of responses to empathy-invoking situations

Affective empathy is commonly measured based on facial, gestural and vocal responses after exposure to an empathy-invoking situation such as someone being hurt (Zhou et al., 2003). Because these measures require trained personnel to recognize such behavioral responses (Losoya & Eisenberg, 2001), they are moderately difficult to be used outside of the research context.

Findings from such situations can also suffer from biases. For instance, the extent of facial, gestural, vocal responses shown can be affected by the extent to which showing such responses was deemed acceptable by themselves, their friends, family and culture, especially concerning negative emotions (Losoya & Eisenberg, 2001). Physiological responses such as heart rate, skin conductance and functional Magnetic Resonance Imaging (fMRI) have been used to overcome some of those challenges of behavioral analyses (Neumann, Chan, Boyle, Wang & Westbury, 2015). However, measuring physiological responses often require specialized equipment, which limits the extent to which they can be used in practice to measure people's level of empathy outside of the research context.

A further limitation of the analysis of responses to empathy-invoking situations is its moderate ecological validity. Ecological validity refers to the extent to which the research setting is similar to the setting participants encounter in their daily lives (Brewer, 2000). While empathy-invoking situations are designed to stimulate what participants face in daily life, the diversity of social, cultural, economic and family background that participants come from make it difficult if not impossible to design a common set of situations that represent such diversity. As a result, such situations might not be able to predict whether participants demonstrate empathy in their daily life.

2.1.2 Performance tasks

On the other hand, both affective and cognitive empathy can be measured on performance tasks such as the Strange Stories Test (Happé, 1994), Sally-Anne False Belief test (Baron-Cohen, Leslie & Frith, 1985), Second Order False Belief test (Baron-Cohen, 1989) and Reading the Mind in the Eyes (Baron-Cohen, Wheelwright, Hill, Raste & Plumb, 2001). For instance, in the Strange Stories Test (Happé, 1994), participants were told of stories involving persuasion, sarcasm or deception. They were then told to determine whether a statement made by a character in the story was true and the reason for their answer. Success in the task depended on the participant's ability to imagine themselves from the perspective of the character. Depending on the performance task, some are easily used out of the research context (Baron-Cohen et al., 2001) while others are moderately easy given the necessity for trained personnel to interpret responses (Happé, 1994).

However, performance tasks of cognitive empathy can also be biased by the use of cognitive strategies to compensate for their lack of intuitive understanding. As an example, some individuals diagnosed with Autism Spectrum Disorder due to their lower levels of cognitive empathy are able to pass performance tests by using complex "if-then" rules (Zalla & Korman, 2018). These rules - such as if the person says something that is unlikely, he/she is likely being sarcastic - allow these individuals to pass the performance test without having cognitive empathy. Performance on these tasks can also be confounded by the socio-cultural background of participants. For instance, certain speech acts (such as sarcasm or humor) might be more acceptable in certain socio-cultural backgrounds (Yue, Jiang, Lu & Hiranandani, 2016), predisposing some participants to a greater sensitivity to them regardless of their empathy.

The limitation of moderate ecological validity from the previous section on empathy-invoking situations also applies to performance tasks.

2.1.3 Self-report questionnaires

Finally, another common method of measuring empathy is through self-report questionnaires. These questionnaires can measure either affective empathy (Spreng, McKinnon, Mar & Levine, 2009), cognitive empathy (Hogan, 1969) or both aspects of empathy (Davis, 1983;

Wakabayashi et al., 2006; Baron-Cohen & Wheelwright, 2004; Jolliffe & Farrington, 2006). These questionnaires typically ask short questions to be answered on a Likert Scale. Because these questionnaires ask questions about a person's experiences in daily life, findings tend to have high ecological validity. For instance, Wakabayashi et al. (2006) included questions of "I often find it difficult to judge if something is rude or polite", "I can pick up quickly if someone says one thing but means another", and "I am good at predicting how someone will feel", which participant can respond on a 4-point Likert Scale.

These questionnaires are generally easy and quick to administer. As an instance, the questionnaire used by Wakabayashi et al. (2006) only has 22 questions that take 2 minutes to answer. Therefore, people outside of the research environment can easily use these questionnaires to find out their level of empathy. Furthermore, such ease of use means that the most established questionnaires such as that of Baron-Cohen and Wheelwright's (2004) derive high generalizability from large samples of over half a million (Greenberg, Warrier, Allison & Baron-Cohen, 2018) and validity in different cultural backgrounds from culturally diverse samples (Wakabayashi, 2013; Dimitrijevic, Hanak, Vukosavljevic Gvozden & Opacic, 2012; Kosonogov, 2014; Groen, Fuermaier, Den Heijer, Tucha & Althaus, 2015).

However, results from self-report questionnaires might be biased. This can result from participants' lack of self-understanding (Wilson & Dunn, 2004). For instance, people might think that they are "good at predicting how someone will feel" even if friends and family around them think otherwise since it might not be socially appropriate for friends and family to inform them of so. The participants' responses might be influenced by their self-efficacy (Caskie, Sutton & Eckhardt, 2014) in their ability to understand others alongside their true empathetic ability. Furthermore, participants might also have self-serving biases (Müller & Moshagen, 2019) such as the social desirability bias (Nederhof, 1985). This means that participants might answer that they are "good at predicting how someone will feel" because this is socially valued rather than truly descriptive of their ability. Together, such biases reduce the validity of findings from self-reported surveys.

2.1.4 Summary

The strength and weaknesses of conventional methods used to measure empathy are summarized in Table 2.1.

	Responses to empathy-invoking situations	Performance tasks	Self-report Questionnaires
Aspects of empathy measured	Affective	Affective and Cognitive	Affective and Cognitive
Biases	Low: Social desirability bias (for non-physiological measures)	Moderate: Sociocultural background; Use of cognitive strategies	High: Limited self-understanding; self-efficacy; self-serving biases
Ecological validity	Moderate	Moderate	High
Ease of being used outside of research context	Low-Moderate	Moderate-High	High

Table 2.1 Strength and weaknesses of conventional methods used to measure empathy

Given the limitations of the conventional approaches used to measure empathy, machine-learning-based analysis of writing was proposed as a novel approach to measure empathy. Empathy can be felt through interpersonal communication (Baron-Cohen & Wheelwright, 2004) and writing is one of the main forms of interpersonal communication. Machine learning can then be used to extract information related to empathy by analyzing statistical patterns in writing (Jurafsky & Martin, 2009).

2.2 Applying machine-learning to measure empathy in writing

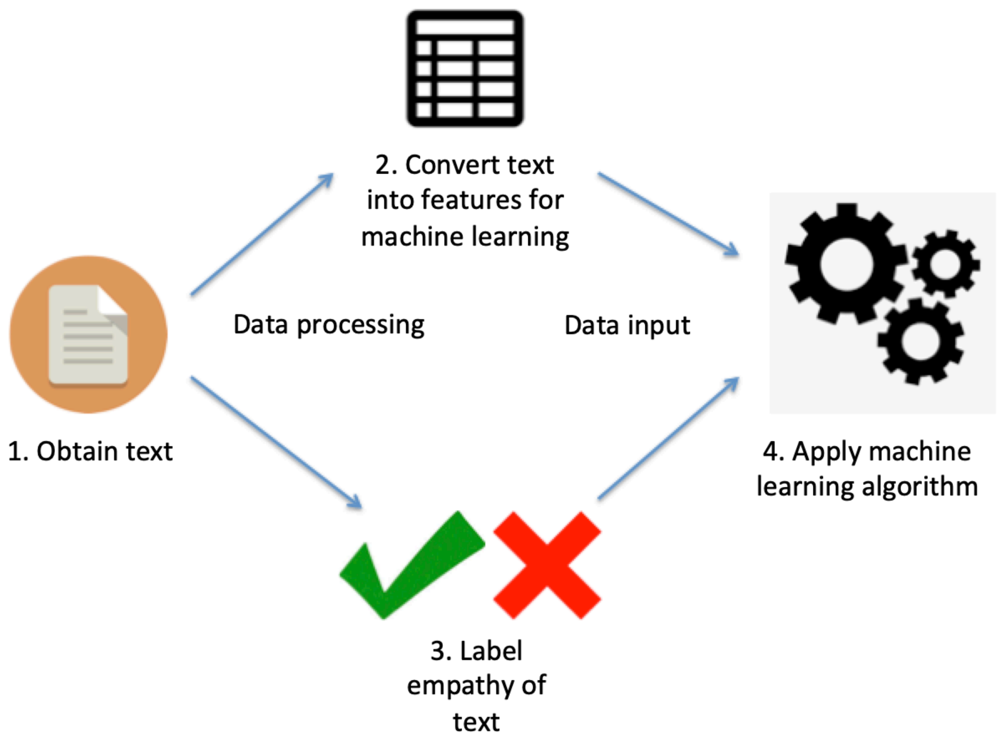

Figure 2.1 General workflow of studies using machine-learning to measure empathy in writing

In this section, studies applying machine learning to measure empathy will be reviewed. The general approach of applying machine learning in each study is shown in Figure 2.1. First, text is obtained. Then text is processed through the steps of labeling its empathy and converting it into features suitable for machine learning. Finally, its labeled empathy and features are used as input into a machine-learning algorithm. More specifically, the machine-learning algorithm uses the features to predict the labeled empathy.

Xiao et al. (2015) and Gibson, Malandrakis, Romero, Atkins & Narayanan (2015) measured the empathy of counselors who conducted counseling sessions for those with alcohol and drug abuse problems. Counseling sessions were taped and converted into text using transcription software. The empathy of each text was labelled by the researchers on an 8-point Likert scale. While Xiao et al. (2015) used words, two-word phrases and three-word phrases as features, Gibson et al. (2015) used words, two-word phrases as well as Linguistic Inquiry and Word Count (LIWC) categories (Tauscizk & Pennebaker, 2010). LIWC is a widely-used software that classifies thousands of words into one of 70 psychology-related categories such as words relating to sensory perception and affect. Finally, the machine learning algorithms of Linear Regression and Naïve Bayes were performed respectively between the features and labelled empathy.

Xiao et al. (2015) found that clarification questions such as “do you think”, “it sounds like” and “you think about” are positively predictive of empathy, supporting how empathy is expressed in counseling situations (Kirschenbaum, 2013). On the other hand, phrases that sound demanding such as “you know what”, “you have to” and “to help you” are negatively predictive of empathy, demonstrating a lack of empathetic concern about the person’s experience (Zickfeld, Schubert, Seibt & Fiske, 2017). Similarly, Gibson et al. (2015) found that “it sounds like”, “sounds” and the words relating to hearing (hear, sound, loud) were positively predictive of empathy. In addition, words relating to perception (loud, touch, darkness) were found to be positively predictive of empathy. This might be explained by people who are better at taking the perspective of others also being more sensitive to perceptual input (Loggia, Mogil & Bushnell, 2008) and therefore write more about them. Finally, words indicating affect (happy, cried, hurt) and anxiety (worried, fearful, nervous) were both found to be positively predictive of empathy, likely because the emotional aspects of empathy can be felt when counselors acknowledge and name emotional states their clients are in (Egan, 2014).

Litvak et al. (2016) conducted a study to quantitatively measure empathy by analyzing writing on social media. Posts and comments were collected from participants’ Facebook accounts and words used by each participant were categorized using LIWC (Tauscizk & Pennebaker, 2010). The writing of each participant was then labeled its empathy based on the participant’s scores in a widely-used empathy self-report questionnaire (Davis, 1983). Finally, participants’ use of words was used to predict their labeled empathy using Poisson regression. This study found that words relating to social processes (mother, rumor, help etc.) were predictive of two aspects of empathy - empathetic concern and perspective taking ($p<0.05$). This was likely because empathy involves understanding other people and being inclined to care for them. Words relating to perception (loud, touch, darkness etc.) were also found to predict perspective taking ($p<0.01$), as found by Gibson et al. (2015).

Khanpour, Caragea and Biyani (2017) measured the empathy of comments on a cancer survivors network forum. Comments were first collected and labeled by the researchers as either empathetic or not. These comments were then transformed into features of words, LIWC categories and advanced word embedding, which represents the syntactic and semantic properties of the word using a long list of numbers (Mikolov, Sutskever, Chen, Corrado & Dean, 2013). Together with the empathy labels, the features were used as input into the machine learning algorithms of Linear Support Vector Classifier, Naïve Bayes as well as advanced neural networks of Convolutional Neural Networks (CNNs) and Long-Short Term Memory networks (LSTMs). While the words and LIWC categories most predictive of empathy were not reported, methods used in this study are novel and informative, especially in comparison with other studies. These will be further discussed in section 2.3 below.

Buechel, Buffone, Slaff, Ungar & Sedoc (2018) measured the empathy of participants after reading sad news stories. Participants were invited to read sad news stories before asking them to write short responses (300-800 characters) that could be shared with their friends. These responses were then transformed into machine-learning features of words and advanced word embedding. Alongside this, they asked people to rate on a 7-point Likert Scale how strongly they felt various emotions related to empathy based on Batson's (1987) Empathetic Concern – Personal Distress Scale. These emotions are warm, tender, sympathetic, softhearted, moved and compassionate. The empathy of each text was then labeled based on the composite score of the strength of various emotions felt by the participant. Finally, the machine learning algorithms of Ridge Regression, Feedforward Neural Network and Convolutional Neural Network were then conducted between the labeled empathy and words. They found that the words most positively predict empathy were "pain", "loss", "heart", echoing the findings of Gibson et al. (2015) concerning words relating to affect. Additional words most predictive of empathy such as "imagine" and "family" can be explained by either a greater ability to consider the perspective of others or its relation to social relationships, as found by Litvak et al. (2016).

2.3 Evaluating studies that apply machine-learning to measure empathy in writing

Methods used in the studies discussed in the previous section will be evaluated in the order of the steps shown in Figure 2.1.

2.3.1 Step 1: Obtaining text

The strength of Litvak et al. (2016) and Khanpour et al. (2007) lies in the use of retrospective writing that is written prior to the research, reducing the risk of participant bias that originates from awareness of the researchers' goals. (McCambridge, Krypi & Elbourne, 2014). The large amount of retrospective writing from social media used also makes the findings more ecologically valid since such writing are based in multiple online social situations, offering an ongoing experiential sample of an individual's online social interactions (Mehl & Conner, 2012) that could reveal their true dispositions (Back et al., 2010). Furthermore, the use of writing done prior to research makes the measurement of empathy outside of the research context to be very easily performed. People would only need to provide access to writing that has been done previously, instead of needing to produce additional writing in order to measure their empathy.

Conversely, Xiao et al. (2015), Gibson et al. (2015) and Buechel et al. (2018) used text written prospectively for the purpose of their studies. Participants might be biased to write in ways that fulfill the purposes of the researchers due to the social desirability bias (McCambridge et al., 2014; Nederhof, 1985). Outside of the research context, participants might not word their responses in similar ways and such findings might not apply. Furthermore, because the response is written for a specific scenario, it is unlikely that empathy presented in these scenarios can generalize to empathy presented in broader settings, leading to poor ecological validity.

2.3.2 Step 2: Converting text into features for machine learning

Text can be converted into features for machine learning based on three approaches. The first approach is to count the number of words in predefined LIWC categories (Tauscizk & Pennebaker, 2010). The second approach is to count the number of times each word occurred. The third approach is an advanced method in which each word is converted into a word embedding that represents the syntactic and semantic properties of the word using a long list of numbers (Mikolov et al., 2013).

Gibson et al. (2015) and Khanpour et al. (2017) found that the second approach far outperformed the first approach. This was likely to occur because each category contained a high number of words (> 800 in some categories). This means that the positive association between individual words and empathy might be concealed by the lack of or negative association between empathy and other words in the same category.

Furthermore, Buechel et al. (2018) and Khanpour et al. (2017) found that the third approach modestly outperformed the second approach. While the third approach demonstrated the best performance in uncovering the association between text and empathy, the method makes it difficult to understand which individual words in the text are responsible for the association between the text and empathy. This means that the findings cannot be compared with those of other studies or discussed in relation to theoretical perspectives. Therefore, the second approach of counting the number of times that each word occurred is optimal for this step.

2.3.3 Step 3: Labeling empathy of text

Two methods - self-reported questionnaires and third-party labeling - are commonly used to label the empathy of text. Both methods are able to measure affective and cognitive aspects of empathy because self-reported questionnaires can contain items on both aspects (Baron-Cohen & Wheelwright, 2004; Wakabayashi et al. 2006) while third-parties make judgments based on both aspects (Gibson et al., 2015, Khanpour et al., 2017; Xiao et al., 2015).

2.3.3.1 Self-reported questionnaires

First, text was labeled empathetic or otherwise based on the writer's empathy as measured on self-reported questionnaires (Buechel et al., 2018; Litvak et al., 2016). Participants were given an empathy score based on their responses to the questionnaire. This score was then used to label the writing of the participant. Due to the dependence on self-reported questionnaires, this method also inherited its limitations of self-understanding, self-efficacy and self-serving biases, elaborated previously in Section 2.1.3.

2.3.3.2 Third-party labeling

Second, writing was labelled as empathetic or not based on researchers who label such writing (Gibson et al., 2015, Khanpour et al., 2017; Xiao et al., 2015). These researchers were third-parties since they were not directly involved in the interpersonal interaction studied. Labels by two researchers for the same set of text were then compared to ensure high interrater reliability. However, because each text had to be manually labeled by a small number of researchers, this method required significant effort from the researchers.

Furthermore, this method of labeling might not accurately capture empathy. This is because empathy concerns understanding a person from their frame of reference (APA, 2020). Third parties who are not engaged in the conversation may not be in the best position to determine whether text are empathetic because third parties do not have information about the frame of reference of the person, which a text is addressed to. For instance, in Khanpour et al. (2017), “Get well soon!” might sound empathetic to the researchers but not to cancer survivors, who might have heard of such platitude often. This is because researchers are likely not cancer survivors, making it difficult for them to accurately imagine the affective and cognitive state that the cancer survivors are in. Instead, the person whom a text is addressed to can most accurately tell if a text is empathetic.

2.3.4 Step 4: Applying machine learning algorithm

Regression methods including Poisson, Linear and Ridge regressions are the most common machine-learning algorithm used (Buechel et al., 2018; Gibson et al., 2015; Litvak et al., 2016; Xiao et al., 2015). This is followed by Naïve Bayes (Khanpour et al., 2017; Gibson et al., 2015) and then Linear Support Vector Classifier (Khanpour et al., 2017). However, none of the algorithms above performs significantly and consistently better.

Advanced neural network algorithms were also used and generally performed better than other algorithms (Buechel et al., 2018; Khanpour et al., 2017). However, they are complex, non-linear algorithms that account for relationships between different words in a text and therefore do not inform which words are individually most predictive of empathy (Karpathy, Johnson & Li, 2015). This means that these words cannot be compared with those from other studies as well as examined from theoretical perspectives.

2.3.5 Overall

Overall, applying machine-learning to measure empathy in writing can have several strengths, as summarized in Table 2.2, when methodological choices are made optimally.

	Machine-learning for analyzing writing
Aspects of empathy measured	Affective and Cognitive
Biases	Depends on how the empathy of text is labeled (step 3)
Ecological validity	High
Ease of being used outside of research context	Very High

Table 2.2 Strength and weaknesses of applying machine-learning methods to measure empathy in writing.

For step 1, this means to use general text that is written prior to a study to improve the ecological validity of its findings and increase the ease of using it outside the research context. For step 2, this means converting text into features for machine-learning by counting the number of times that each word occurred. This approach balances performance of the machine-learning method and the ease of discussing its findings with those from other studies and theoretical perspectives. For step 4, this means to experiment with Regression, Naïve Bayes and Linear Support Vector Classifier. These machine-learning algorithms perform equally well and produce findings that can be compared with results of other studies and examined based on theoretical perspectives on empathy.

	Self-report Questionnaires	Third-party labeling
Biases	High: self-understanding, self-efficacy and self-serving biases	High: Third-party may not appreciate what is empathetic accurately

Table 2.3 Comparison of the bias of methods used to label empathy of writing

However, there is no optimal methodological choice available for step 3, meaning that labeling empathy of text in step 3 forms a critical bottleneck. Both existing solutions to labeling empathy of text in step 3 contain high biases, as shown in Table 2.3, Self-report questionnaires are limited by self-understanding, self-efficacy and self-serving biases while third-party labeling is restricted by the inability of third-parties to appreciate what is empathetic accurately. Therefore, an alternative method of labeling the empathy of text is necessary to realize the potential of applying machine-learning methods to measure empathy in writing.

3 The Present Study

The present study aims to explore the validity of a new method of labeling empathy of text in order to improve machine-learning methods of measuring empathy in writing. In particular, the new method uses the effective helpfulness of people's written help (rated by the recipient of the written help) to label the empathy of writing. This method was devised to attempt to overcome some limitations identified in current methods in 2 ways. First, it overcomes the self-understanding, self-efficacy and self-serving biases associated with self-reported questionnaires since the empathy of individuals is not labeled based on their own opinions. Second, it overcomes the inaccurate labeling of empathy in text by third-parties since empathy is labeled by the recipients of written help (second-parties) who can accurately judge empathy.

3.1 Research questions

RQ 1. Is the effective helpfulness of written help, as rated by the recipient of the help, associated with the level of empathy of helpers?

RQ 2. Which words are significant predictors of the effective helpfulness of people's written help?

RQ 3. Which words used generally by people in social media writing are significant predictors of their levels of effective helpfulness?

RQ 1 seeks to investigate whether the effective helpfulness of people's written help to others is a valid method of labeling empathy of writing, based on its association with their level of empathy. RQs 2 and 3 seek to reveal the words that machine-learning methods find to be significant predictors of effective helpfulness, as a measure of empathy. RQ 2 focuses on the words used in giving written help within a help-seeking forum while RQ 3 focuses on words used on social media (Reddit) more generally to improve the generalizability of findings.

4 Methods

4.1 Measures

The effective helpfulness of written help is measured both at the level of a single written help and the level of the person who has given written help multiple times. The empathy of a person is also measured using the short form of the Empathy Quotient questionnaire.

4.1.1 Effective helpfulness of written help (comments)

The study was conducted using written text on Reddit. Reddit is a widely used online forum, which is the 4th and 19th most visited website in the United Kingdom and the World respectively (Alexa, 2020). Reddit is structured into forums (also known as "subreddits") that are organized around user communities with common interests. Social media platforms including Reddit are ecologically valid methods of collecting writing (Back et al., 2010; Mehl & Conner, 2013) because such writing is produced naturally by users, rather than prompted by experimenters.

The Advice forum, where most of the comments are taken from, is a community where users ask for advice on issues they encounter in daily life (Reddit, 2020a). These issues include problems with family and friends, difficulties at school/work as well as troubles in pursuing one's interests and hobbies. Under the posts, users can provide written help to the post author by commenting. In the Advice forum, post authors can label the comment(s) that they have found helpful, similar to liking a comment on Facebook. This can indicate which comment(s) the post authors have found to be able to understand and address the concerns that they have raised.

Comments that were not labeled as helpful were deemed to be unhelpful. To minimize mislabeling of unhelpful comments, comments on posts with no helpful comment were excluded. This was done since authors of those posts likely did not actively label comments based on whether they were helpful.

An example is presented below to illustrate the difference between helpful and unhelpful comments.

Help-seeking post or question

> My friend is upset and she won’t tell me why, I asked what I can do to help her out and she says she doesn’t want me to do anything. What do I say/do?

Helpful comment

> Give her space but let her know you respect her request. Say that when and if she wants to share or even if she just wants a distraction, you're there to help. No judgments, she's your friend no matter what.

Unhelpful comment

> You ignore it until she is grown up enough to communicate. This is an age old worn-out game played by people who can't communicate properly. Don't play the game

4.1.2 Effective helpfulness of users

Building on the approach above, the effective helpfulness of each user (proportion of helpful comments) was found using Equation 4.1. Proportion rather than the raw number of helpful comments was used to control for the difference in the number of comments made by each person.

$$\text{Effective helpfulness of a user } = \frac{No.of\ helpful\ comments}{No.of\ comments}$$

Equation 4.1 Effective helpfulness of a user (Proportion of helpful comments)

Because empathy measures the extent to which a person can understand and adequately respond to others (Baron-Cohen & Wheelwright, 2004), the proportion of time that a person can be helpful to others can serve as a good measure of empathy. This measure captures both the affective and cognitive aspects of empathy because both are necessary in responding adequately to help-seeking posts. The appraisal of empathy by people who the help comments are directed to can most accurately determine whether an individual has managed to understand them from their frame of reference, which is at the core of the definition of empathy (APA, 2020). Such a conception of empathy is built on the ontological tradition of interpretivism (Brundrett & Rhodes, 2013) since whether help comments are considered to be helpful depends on how they are interpreted by their intended recipients.

Furthermore, to ensure that the measure assesses empathy, it is necessary to explore the relationship between this measure and constructs related to empathy. For example, effective helpfulness could be confounded with people's level of pro-sociality. Pro-social behavior refers to behavior that intends to benefit other people or society as a whole (Eisenberg, Fabes & Spinrad, 2007; Jensen, 2016). Pro-social behavior is more related to how often a person attempts to help others rather than their frequency of effectively helping others by understanding their issue and responding appropriately (Davis, 2015)

Effective helpfulness could also be confounded with emotional intelligence and interpersonal intelligence. Based on Goleman (1998), emotional intelligence involves self-awareness, self-regulation, motivation, empathy and social skills. While termed differently, Gardner's (1993) concept of interpersonal intelligence is highly overlapping, concerning sensitivity to others' moods, feelings, temperaments, motivations as well as abilities to empathize, communicate and cooperate with others. Among the skills that constitute emotional intelligence/interpersonal intelligence, effective helpfulness mostly indicates about a person's empathy. This is because effective helpfulness may not be related to self-awareness, self-regulation, motivation as well as aspects of social skills that do not tap on empathy such as the ability to initiate social interactions.

Overall, the measure could be argued to be mostly representative of empathy.

4.1.3 Short form of Empathy Quotient questionnaire

The short form of Empathy Quotient questionnaire (Wakabayashi et al. 2006) is an established measure of empathy with over 300 citations on Google Scholar. Items came from the long form of Empathy Quotient questionnaire (Baron-Cohen & Wheelwright, 2004) with over 3,500 citations on Google Scholar. The short form was chosen to reduce the time taken to answer the questionnaire and increase the response rate. The short form is a 22-item forced-choice self-report questionnaire that can be answered on a four-point Likert Scale (Strongly Agree, Agree, Disagree, Strongly Disagree). Questions include "I often find it difficult to judge if something is rude or polite", "I can pick up quickly if someone says one thing but means another", and "I am good at predicting how someone will feel".

Each response can give 0, 1 or 2 points, leading to a maximum total score of 44. A higher score represents greater empathy. Items on the questionnaire cover both affective and cognitive aspects of empathy, providing high content validity. It has high internal consistency (Cronbach's alpha=0.90) and test-retest reliability after 12 months ($r = 0.97$, $p < .001$) (Wakabayashi et al. 2006). Low performance on this questionnaire also predicts a higher likelihood of being diagnosed with Autism Spectrum Disorder, a group with known deficits in empathy (Wheelwright et al., 2006). While this study acknowledges the limitations of this questionnaire, as discussed in section 2.1.3, this questionnaire is used as an initial tool to explore the validity of the effective helpfulness as a method for labeling empathetic writing. The questionnaire is available in appendix A.

4.2 Participants and Data

4.2.1 Effective helpfulness of written help (comments)

Text from Reddit was downloaded through the Pushshift Application Programming Interface (Pushshift, 2020). Suitable posts and all associated comments from the Advice subreddit were downloaded in the past 300 days. Among the 24964 posts that were downloaded, there were 92477 comments (41146 helpful). Comments by the post authors and automated bots were excluded.

4.2.2 Effective helpfulness of users

Furthermore, posts and comments within Reddit (not just the Advice subreddit) of 508 users, each with more than 20 comments in the Advice subreddit, were downloaded. This aided to understand which words found in users´ general engagement within Reddit (not just those targeted at providing help comment within the Advice forum) were predictors of their effective helpfulness.

4.2.3 Short form of Empathy Quotient questionnaire

508 Reddit users with more than 20 comments in the dataset were sent an online questionnaire through Reddit and 91 responded. Gender and age were optional to report. 86 participants reported gender (53 male and 33 female) and 83 reported age (M=33.7, SD=13.8).

4.3 Data preparation

4.3.1 Effective helpfulness of written help (comments)

Only one comment from each author and in each post was retained. Details of this process are available in appendix B. This was done to ensure comments are independent from one another in order to fulfill an assumption of logistic regression, which is subsequently performed. As a result, there remain 10607 comments (6020 helpful) containing 4397 unique words. On average, each comment has 105 words (SD=127).

4.3.2 Effective helpfulness of users

Only users with more than twenty comments were included to minimize the likelihood that their effective helpfulness was biased due to chance events. Because 508 selected users have over 3 million posts and comments, the size of the data (over 3 million samples with 8734 independent-variables/unique-words) made it impossible to process such data without specialized computing hardware. Therefore, 100 posts/comments were randomly sampled from each user to represent the writing of each user. Each post/comment was labeled with the proportion of helpful comments in the Advice forum of its author. Among the 50800 posts/comments, each post/comment has 44.6 words (SD=77.8). The mean effective helpfulness of users is 0.5440 (SD=0.1956).

4.3.3 Machine learning

Text was processed based on counting the number of times each word occurred, based on best practices in Literature Review section 2.3.2. First, text was split up into individual words. The number of times each word occurred in each text was then counted. Words that occurred fewer than ten times altogether were removed to minimize the effects of misspelled or rare words.

4.4 Statistical analysis

A linear regression was conducted between the proportion of helpful comments and Empathy quotient score (Wakabayashi et al., 2006). Assumptions of the model were checked and largely satisfied with details in appendix C.

Using a two-sample t-test, this distribution of EQ scores (M=24.45, SD=8.822, N=91) in this study is found to be not significantly different (t(1850)=0.0169, two-tailed p=0.9866) from the sample (M=23.8, SD=8.75, N=1761) in Wakabayashi et al. (2006), demonstrating the representativeness of the sample in this study.

4.5 Machine learning methods

4.5.1 Effective helpfulness of comments and word use

4.5.1.1 Choosing a machine learning algorithm

To determine the optimal machine-learning algorithm for identifying words that predict the effective helpfulness of a comment, suitable algorithms reviewed in Literature Review section 2.3.4 were evaluated. A popular metric in machine learning, accuracy (Jurafsky & Martin, 2009) was used as to evaluate the algorithms. The accuracy was calculated based on the proportion of text
that were correctly predicted its effective helpfulness by each algorithm.
Table 4.1 shows the accuracy of each model. Because of its good performance, Logistic Regression was chosen as the algorithm to be used subsequently.

Model	Accuracy
Logistic Regression	84%
Linear Support Vector Classifier	82%
Multinomial Naïve Bayes	79%

Table 4.1 Accuracy of machine learning algorithms in predicting effective helpfulness of comments.

4.5.1.2 Logistic Regression

Logistic Regression seeks to estimate perimeters (β) for the logistic function in Equation 4.2 using Maximum Likelihood Estimation (MLE) (Bishop 2006).

$$log\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{n}\beta_i * Word_i$$

$$where\ \text{odds} = \frac{p}{1-p}\ and\ \ MLE\ maximizes\ \prod_{i=1}^{m} p_i{}^{y_i} * (1-p_i)^{1-y_i}$$

Equation 4.2 Logistic Regression p is the predicted effective helpfulness of the comment (a value between 0=unhelpful and 1=helpful), β_0 is the constant term, β_i is the beta of $\square ord_i$, $Word_i$ is the number of times that a particular word occurs in the comment, n is the number of unique words, m is the number of comments and y_i is the true effective helpfulness of a comment (0=unhelpful, 1=helpful). All β are unstandardized.

Assumptions of the model were checked and largely satisfied with details in appendix C.

4.5.2 Effective helpfulness of users and word use

4.5.2.1 Linear regression

A different algorithm had to be used to identify words that significantly predict effective helpfulness of users since it is a proportion between 0 (unhelpful) and 1 (helpful) while the effective helpfulness of comments is either 0 or 1. Based on Literature Review section 2.3.4 , linear regression was used to identify words in the general Reddit writing of analyzed users that predict their effective helpfulness. Linear regression seeks to estimate perimeters (β) for Equation 4.3 using Mean Squared Error (MSE) to minimize the error term (Bishop, 2006).

The choice of linear regression was a compromise because algorithms that are guaranteed to predict the effective helpfulness of each user within 0 and 1 such as Fractional Logit regression (Statsmodels, 2020a) were not optimized for a large number of (> 5000) words as independent variables and therefore could not run without specialized computing resources.

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i * Word_i + \varepsilon$$

$$MSE\ minimizes\ \frac{1}{m} \sum_{i=1}^{m} \varepsilon^2$$

Equation 4.3 Linear Regression. y is the proportion of helpful comments that a user has. β_0 is the constant term, β_i is the beta of $Word_i$, $Word_i$ is the number of times that a particular word was used by the user, n is the number of unique words, , m is the number of posts/comments, and ε is the error term. All β are unstandardized.

Assumptions of the model were checked and largely satisfied with details in appendix C.

4.6 Ethics

This research has been granted ethical clearance from the R&I ethics committee, Faculty of Education, University of Cambridge. Ethical guidelines given by the British Educational Research Association [BERA] (2018) and The British Psychological Society [BPS] (2014; 2018) were followed in this research. Details of ethical challenges and solutions to these challenges are found in Appendix D.

5 Results

5.1 Research Question 1

RQ 1. Is the effective helpfulness of written help, as rated by the recipient of the help, associated with the level of empathy of helpers?

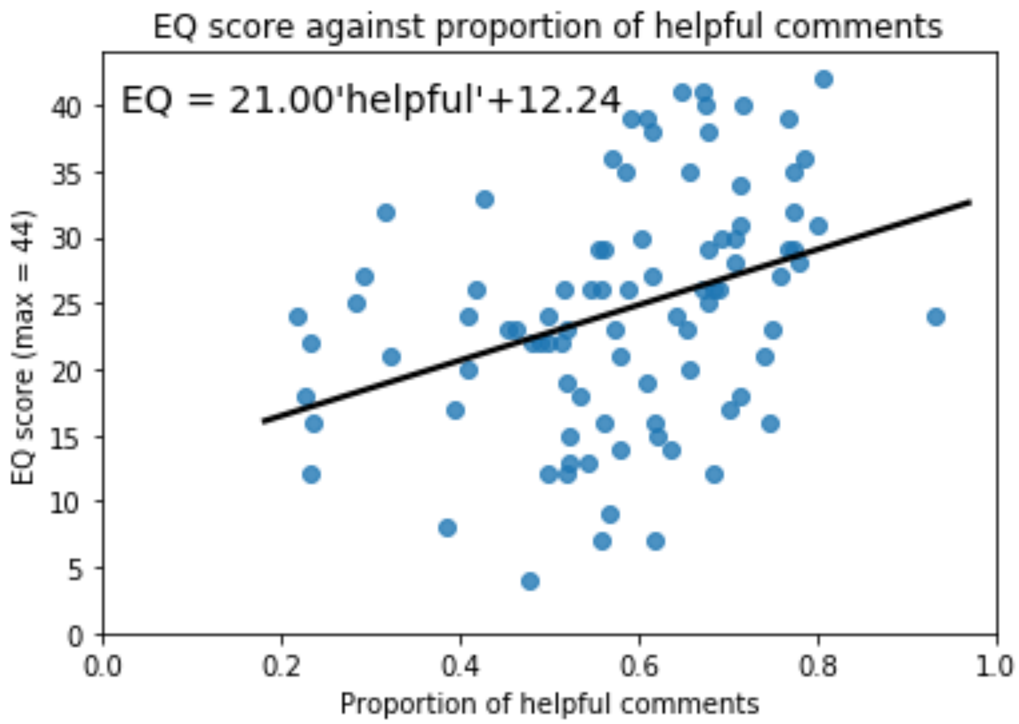

Figure 5.1 Empathy quotient (EQ) score against proportion of helpful comments. EQ refers to Empathy quotient (EQ) score of users while 'helpful' represents the effective helpfulness of users (proportion of helpful comments).

As illustrated in Figure 5.1, the regression equation is EQ = 21.00 'helpful' + 12.24 ($r(91)=0.359$, two-tailed $p < 0.001$) with a moderate correlation effect (Cohen, 1992). Without making the assumptions of linear regression, the non-parametric Spearman's Rank Correlation $r_s=0.4235$, N=91, two-tailed $p < 0.001$. The moderate correlation effect demonstrates that EQ and effective helpfulness (proportion of helpful comments) are likely to be measuring the same underlying construct of empathy to some degree. This shows that the measure of effective helpfulness (proportion of helpful comments) has moderate convergent validity with EQ scores, a well-established measure of empathy.

5.2 Research Question 2

RQ 2. Which words are significant predictors of the effective helpfulness of people's written help?

Using logistic regression, words in written help comments were used to predict their effective helpfulness. Because logistic regression does not have an R-squared value, two alternative statistics – F1 score and Area Under Curve of Receiver Operator Characteristic Curve - were used to measure the goodness of fit (Jurafsky & Martin, 2009).

F1 score

As shown in Equation 5.1, F1 score is a composite score of precision and recall. Precision refers to the proportion of correctly predicted comments out of comments predicted to belong to a class. In contrast, recall refers to the proportion of correctly predicted comments out of comments that belong to a class.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Equation 5.1 Formulae for F1 score. The first word (True/False) indicates if the prediction was made accurately while the second word (Positive/Negative) indicates the true effective helpfulness of the comment (Positive=Helpful, Negative=Unhelpful).

True effective helpfulness	n	Precision	Recall	F1-score

Unhelpful	4585	0.80	0.83	0.82
Helpful	6108	0.87	0.84	0.86
Total	10693	0.84	0.84	0.84

Table 5.1 F1 score of logistic regression between the words in help comments and effective helpfulness.

The high F1-score of helpful and unhelpful comments (0.82 and 0.86) in Table 5.1 demonstrates good performance of the model across both helpful and unhelpful comments.

Area Under Curve of Receiver Operating Characteristic Curve (ROC AUC)

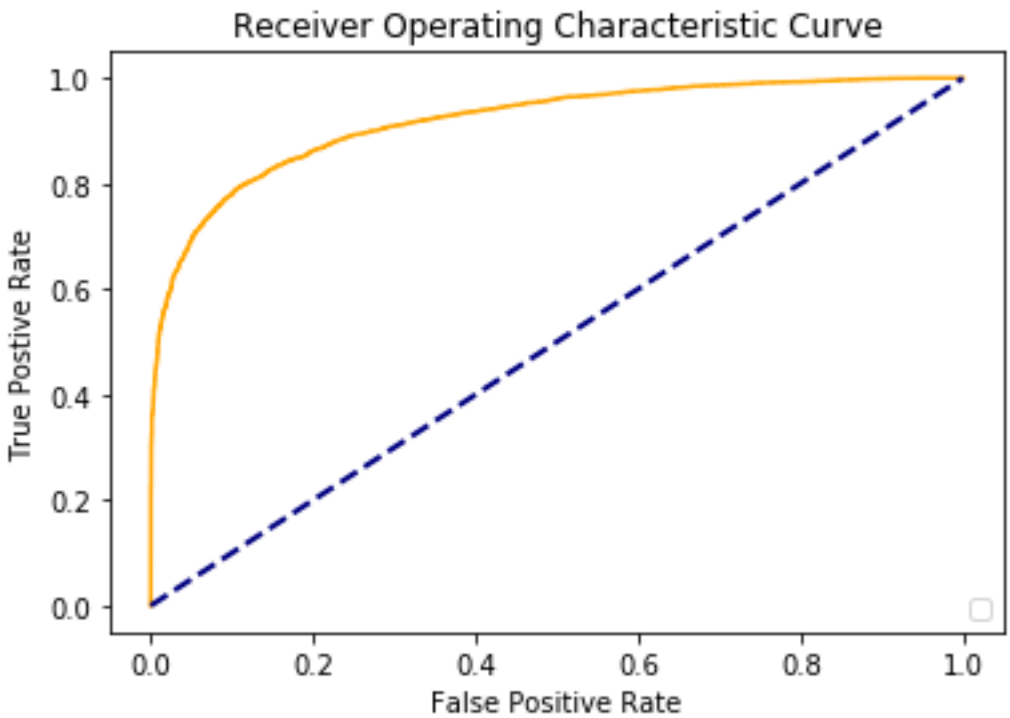

Figure 5.2 Receiver Operating Characteristic Curve

Area Under Curve of Receiver Operating Characteristic Curve (ROC AUC) was used as an alternative measure of goodness of fit. As shown in Figure 5.2, the Receiver Operating Characteristic Curve plots the True Positive Rate (TPR) – the proportion of correct predictions - against the False Positive Rate (FPR) – the proportion of incorrect predictions -

at all decision boundaries of the logistic regression classifier. A good classifier can achieve high TPR while keeping FPR low. Therefore the closer the TPR is to one at all levels of FPR, the better the classifier. The extent to which a classifier can achieve this is represented by ROC AUC, which is found by integrating the curve with respect to the x-axis. ROC AUC=0.9182. This means that for a value that ranges from 0 to 1, the performance is good.

50 and 44 words are found to be positive and negative significant predictors of effective helpfulness respectively (one-tailed $p < 0.05$). Word clouds are presented in Figures 5.3 and 5.4 as graphical representations of the relative significance of each word in predicting effective helpfulness of written help comments. Furthermore, meaningful class themes were found based on similarities between these words. Statistical details of words are available in Tables E.1 and E.2 of appendix E.

5.2.1 Significant positive predictors of effective helpfulness of comments

Figure 5.3 Word cloud of significant positive predictors of effective helpfulness of a comment

Class name	Words	Examples
Polite, friendly-sounding words	glad, welcome, personally, friend, feels, hey	--> Me, **personally**...I’d let it slide. He’d be --> That's okay I'm just **glad** that you were able to --> maybe text her? Be like **hey**, just wanted to say thanks

Optimistic sounding words	luck, hopefully, yes, hope, forward, strong, good, learned, helped,	--> session with your therapist. Good **luck** I am going to give --> **hope** something I say can help you a little! --> back off my life. And **yes** that is dangerous and quite --> work that you can look **forward** to. Gaming is awesome but
Words addressing the post author directly	you	--> But I really think **you** deserve better. **You** sound like --> I understand that **you** really like these guys --> as long as **you** feel **you** are making the best of

Table 5.2 Class themes of significant positive predictors of effective helpfulness of a comment and examples

5.2.2 Significant negative predictors of effective helpfulness of comments

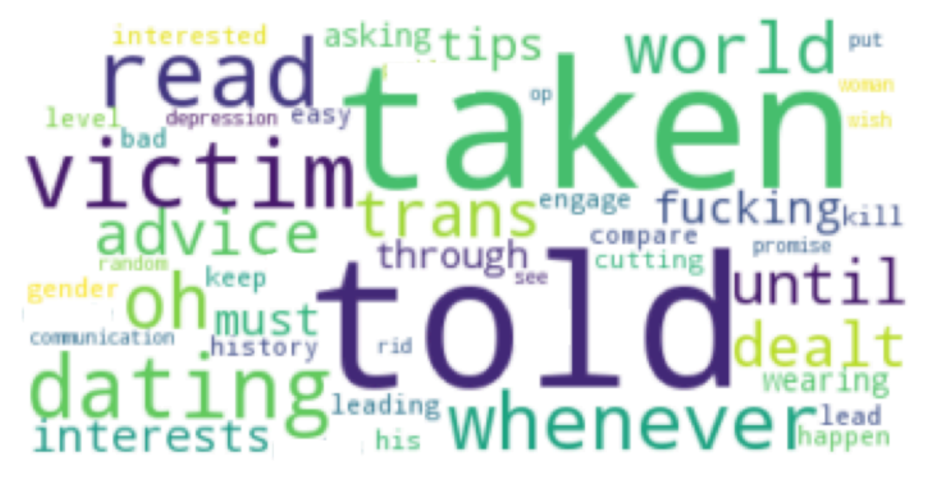

Figure 5.4 Word cloud of significant negative predictors of effective helpfulness of a comment

Class name	Words	Examples
Words indicating negative emotions	bad, depression, rid, victim, fucking, kill	--> to be labelled as a "**victim**". She might be afraid of --> actually pretty fun, but really **fucking** sucks when the racism breaks

		--> because I was internalizing every **bad** thing that happened, and bottling
Words that trivialize the problem faced by the post author	easy, advice, told, dealt, wish, promise	--> it's the latter, as I **dealt** with when I was like --> because it seems like the **easy** solution to your situation. --> The best **advice** I can give you though

Table 5.3 Class themes of significant negative predictors of effective helpfulness of a comment with examples.

5.2.3 Overall

To answer RQ 2, polite, friendly-sounding words, optimistic-sounding words and words addressing the post author directly significantly and positively predict the effective helpfulness of people's written help comments. This is shown in Table 5.2. On the other hand, words that indicate negative emotions and trivialize the problem faced by the post author significantly and negatively predict the effective helpfulness of people's written help comments. This is shown in Table 5.3.

5.3 Research Question 3

RQ 3. Which words used generally by people in social media writing are significant predictors of their levels of effective helpfulness?

Using linear regression, word use by users of Reddit was found to significantly predict their effective helpfulness ($F(8734, 50800)=6.950$, two-tailed $p<0.001$, $R^2=0.5903$).

More than 50 words are found to be positive and negative significant predictors of effective helpfulness respectively (one-tailed $p<0.05$). Therefore, 50 words that are most significant predictors of effective helpfulness, both positively and negatively are presented. Word clouds are presented in Figures 5.5 and 5.6 as graphical representations of the relative significance of each word in predicting effective helpfulness of users. Statistical details of these words are available in Tables E.3 and E.4 of appendix E.

5.3.1 Significant positive predictors of effective helpfulness of users

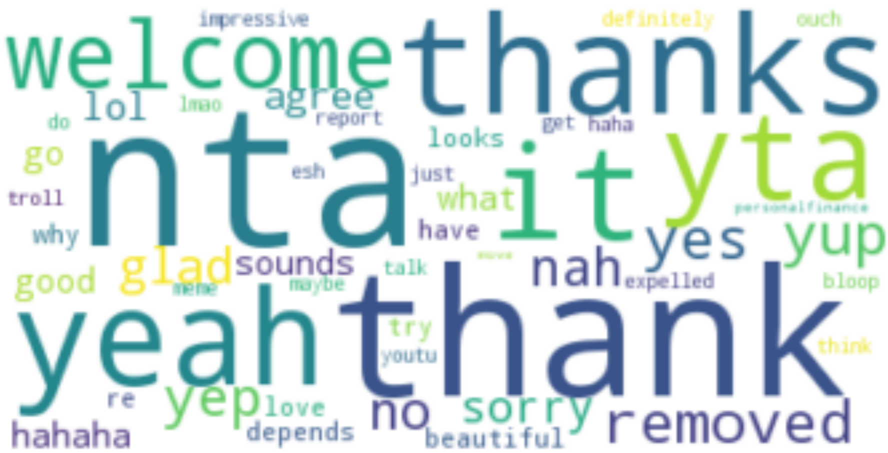

Figure 5.5 Word cloud of 50 most significant positive predictors of effective helpfulness of a social media (Reddit) user

Class name	Words	Examples
Polite friendly-sounding words	thanks, thank, yeah, welcome, yup, yes, yep, sorry, lol (laughing out loud), agree, hahaha, haha, lmao (laughing my ass off), ouch	--> almost had a heart attack **haha**. **Thanks** for the help! --> back at now and say "**yeah**, that was pretty good"? --> I **agree** with you 100 percent. I --> The answer is **yes**. Following your passion is great
Optimistic sounding words	good, beautiful, love, glad, impressive	--> yourself motivated each day and **good** things will arrive --> You're a **beautiful** person, remember --> that was pretty damn **impressive**!
Question words	why, what	--> Could you explain **why** you've been posting this --> turn into a career? - **what** were your favorite classes in

Table 5.4 Class themes of 50 most significant positive predictors of effective helpfulness of a social media (Reddit) user with examples. Explanation of abbreviations in parenthesis.

5.3.2 Significant negative predictors of effective helpfulness of users

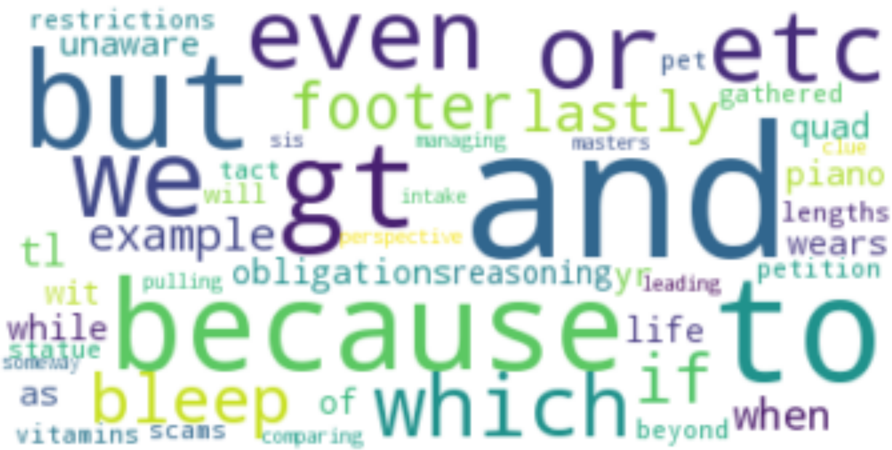

Figure 5.6 Word cloud of 50 most significant negative predictors of effective helpfulness of a social media (Reddit) user

Class name	Words	Examples
Conjunctions (suggesting longer sentences)	And, but, to, because, if, while, which, when, beyond	--> **and** you will become stronger **and** the next bully will have --> think the OP feels better **because** they now have options, **which** --> together in a small space **but** instead of talking **to** each
Words indicating cognition	reasoning, wit, tact	--> her business, and explain your **reasoning**. If you can, wait to --> combined with their token quick **wit**, I find it hysterical, but --> that would be a good **tact** to take in your response.
Words indicating conflict	restrictions, scam, petition, pulling, obligations	--> government puts more and more **restrictions** on the products I enjoy. --> first, and would need a **petition** for emancipation of a minor --> I've seen numerous **scam** emails on my work email

Table 5.5 Class themes of 50 most significant negative predictors of effective helpfulness of a social media (Reddit) user with examples.

5.3.3 Overall

To answer RQ 3, polite, friendly-sounding words, optimistic-sounding words and question words most significantly and positively predict the effective helpfulness of a social media (Reddit) user. This is shown in Table 5.4. On the other hand, conjunctions, words indicating cognition and words indicating conflict most significantly and negatively predict the effective helpfulness of a social media (Reddit) user. This is shown in Table 5.5.

6 Discussion

The aim of this study was to explore the validity of a new method of labeling empathy of text in order to improve machine-learning methods of measuring empathy in writing. To achieve this, three research questions were posed. RQ1 explored if the effective helpfulness of people's help comments is associated with their level of empathy. RQ 2 and 3 explored how machine-learning methods can use words contained in help comments (RQ 2) and general engagements on a popular social media platform, Reddit, (RQ 3) to predict effective helpfulness as measured above. In summary, results indicate that first, the effective helpfulness is a valid method of labeling empathy of text and, second, words that are positive and friendly, attempting to understand the perspective of others and indicative of an absence of elaborate argumentation are predictive of effective helpfulness.

6.1 Research Question 1

Beyond theoretical justifications for the construct, discriminant and ecological validity of effective helpfulness as a measure of empathy, results show that effective helpfulness has moderate convergent validity with an established measure of empathy, the short form of the Empathy Quotient questionnaire (Wakabayashi et al., 2006). Therefore, the naturalistic labeling of effective helpfulness may be a novel and suitable method to label the empathy of writing for machine-learning analysis of writing.

6.2 Research Questions 2 and 3

Significant predictors of effective helpfulness in comments in Reddit Advice forum and general social media (Reddit) users overlap in polite, friendly-sounding words and optimistic-sounding words. Furthermore, class themes in comments and users are related to each other. For instance, both words addressing the post author directly in comments and question words in users suggest an attempt to understand the perspective of others. Therefore, subsequent analysis of class themes that significantly predict effective helpfulness in both comments and users will be combined.

Three overarching themes that emerge as significant predictors of effective helpfulness are words that are positive and friendly, attempting to understand the perspective of others and indicating an absence of elaborate argumentation.

The first overarching theme is positive and friendly words. Effective helpfulness, as a measure of empathy, is positively predicted by polite friendly-sounding and optimistic-sounding words but negatively predicted by words that indicate negative emotions or interpersonal conflict. This supports the positive correlation that has been found between empathy and the attributes of friendliness (Melchers et al., 2016) and optimism (Hojat, Vergare, Isenberg, Cohen & Spandorfer, 2015).

This overarching theme has been found in previous literature on applying machine learning to measure empathy in writing. Affect-related words (such as happy, sad, tears) were found to be positively predictive of empathy in counselors (Gibson et al., 2015). While Gibson et al. (2015) considered words indicating positive and negative affect together, this study revealed that words relating to positive affect and negative affect were positively and negatively predictive of empathy respectively. Similarly, Litvak et al. (2016) found that empathy is positively predicted by words relating to social processes (mother, rumor, help etc.) but did not differentiate between words indicating supportive and conflicting social processes. This suggests that the approach of investigating individual words in this study is able to reveal more nuanced associations between words and empathy compared to investigating word categories, as in Gibson et al. (2015) and Litvak et al. (2016).

However, findings from other literature contradict some aspects of this overarching theme. For instance, Buechel et al. (2018) found that "pain", "loss" and "terrible" were found to be positively predictive of empathy whereas this study found words relating to negative emotions ("depression", "rid" and "bad") to be negatively predictive of empathy. This is likely because empathy was measured in Buechel et al. (2018) based on participants self-reporting the intensive of various emotions related to empathy. This meant empathy was operationalized to only include the affective, but not cognitive aspects. As a result, the words in Buechel et al. (2018) are most predictive of whether participants think they are affectively empathetic.

On the other hand, the measure of effective helpfulness in this study operationalizes both cognitive and affective aspects of empathy because both are necessary in responding adequately to help-seeking posts. Furthermore, because when using effective helpfulness, empathy is judged from the perspective of the help-seeking post author, it more accurately captures whether the commenter is able to understand others from their frame of reference, which is at the core of the definition of empathy (APA, 2020). Finally, writing in Buechel et al. (2018) consists of responses to sad news stories while writing in this study consists of comments to help-seeking posts. This difference in the setting of the writing might also have contributed to such contradiction given that the effective display of empathy depends on the circumstance (Hoffman, 2000).

Another overarching theme is words that suggest attempts to understand the perspective of others. Helpful people do so by asking more questions and addressing other people directly, instead of trivializing the difficulties faced by others. This supports existing literature on the association between empathy and attempts to understand the viewpoints of others (Baron-Cohen & Wheelwright, 2004; Weinera & Austerb, 2007). Terms indicating an inclination to find out more about the perspective of others were also found by Xiao et al. (2015) to predict empathy, even though different terms of "do you think", "it sounds like" and "you think about" were found.

The final overarching theme that predicts effective helpfulness is the absence of elaborate argumentation. Elaborate argumentation is presented through words indicating cognition as well as the use of conjunctions that suggests longer sentences. Language involving elaborate argumentation may not be seen as empathetic because such language is less able to provide emotional solace, a critical aspect of affective empathy (Burleson, 2009; Jones & Wirtz. 2006). However, elaborate argumentation only predicts effective helpfulness in social media (Reddit) users generally instead of specifically in written help comments on the Advice forum within Reddit. This suggests that empathy is presented using different words when providing help-comments compared to engaging on social media (Reddit) generally. The context-specific nature of empathetic words supports the understanding that whether a behavior is seen as empathetic depends on the situation in which it is carried out (Hoffman, 2000).

6.3 Limitations

This study has some methodological limitations that are important to acknowledge.

The machine learning method applied was solely based on the word count (the number of times each word occurred). Therefore, the method did not reveal possible predictors of empathy that required nuanced reading of writing in context. For instance, some words (such as wears) have multiple meanings depending on the context, as shown in Table 6.1. The various meanings of these words may not predict empathy similarly. For example, the expression "the shock wears off" is more empathic than "wears a short sleeve shirt". However, the current method is not able to consider such differences in meaning.

Effective helpfulness	Excerpts of comments
Unhelpful	--> neck or arms if he **wears** a short-sleeve shirt.
Helpful	--> feel worse as the shock **wears** off. You're alone (I believe?)

Table 6.1 Comments mentioning the word "wears"

Furthermore, even when words are used to express similar meaning, the tone of the writing might express empathy differently. In Table 6.2, words most predictive of unhelpful comments and of helpful comments are similar in terms of their intention to assist others. However, unhelpful comments word their attempts to assist in a demanding, instructional tone whereas helpful comments do so in a gentle, recommendatory tone. Such differences in tone are not accounted for in the machine learning approach used in this study.

Effective helpfulness	Excerpts of comments
Unhelpful	--> My **advice**. DONT GET MARRIED.
	--> This is what therapy can **help** you with, they'll give you
Helpful	--> But my **advice** is it won't hurt to
	--> hope something I say can **help** you a little!

Table 6.2 Comments demonstrating the importance of context. Words most predictive of effective helpfulness in bold.

To overcome the problem of words with multiple meanings, multi-word phrases involving two or three words can be considered in place of single words to consider the context in which they are used. Such an approach was used by Xiao et al. (2015), which found phrases such as “it sounds like” and “do you think” to be most predictive of empathy. In using multi-word phrases for machine learning, caution however needs to be taken since the greater number of possible multi-word phrases means that the number of times each multi-word phrase occurs is lower. For instance, if “wears” occurs twice, “wears a” and “wears off” may only occur once each. This might reduce the statistical significance of each multi-word phrase as a predictor. To avoid such a situation, the use of a larger text corpus and correspondingly specialized computing hardware might be considered.

Also, caution needs to be taken in relation to the generalizability of this study’s findings. The study measured empathy in writing from Reddit, an online forum. Because writing in different contexts is likely to be different, the identified words most predictive of empathy might not be applicable outside of the context of online forums and social media. Furthermore, the writing is not representative of people from various demographics. Among the participants of this study, 62% are males with an average age of 33.7 (SD=13.8). Furthermore, Reddit users are overwhelmingly represented by people living in Canada, the United Kingdom and the United States of America (Alexa, 2020). This suggests that the findings are unlikely to apply to demographics significantly different from this group since demographics influence language use (Nguyen, Seza Doğruöz, Rosé & de Jong, 2016).

6.4 Future work

While the results of this study are promising, more work needs to be done to validate effective helpfulness as a measure of empathy. To do so, the relationship between the effective helpfulness of users can be compared to more forms of established measures of empathy such as other self-reported questionnaires (Jolliffe & Farrington, 2006), performance tasks (Baron-Cohen et al., 2001) and responses to empathy-invoking situations (Losoya & Eisenberg, 2001). Empirical measures of constructs related to empathy such as pro-sociality (Eisenberg et al., 2007) and emotional intelligence (Goleman, 1998) can also be used to confirm the discriminant validity of effective helpfulness as a measure of empathy.

Furthermore, words can be used to predict users' self-reported levels of empathy directly to determine the extent to which they are similar to words that predict effective helpfulness. Finally, the method used in this study can be applied to study writing from other sources and demographics to derive findings that apply to them. When language data is only available in the form of speech, speech can be first automatically transcribed (Xiao et al., 2015) before applying a similar method. Together, it means this method can facilitate the study of empathy in a wide range of circumstances involving interpersonal language use.

7 Conclusion

Overall, this study addresses a critical gap in the research around the measurement of empathy. The study contributes by proposing and validating a new way to apply machine-learning methods to measure empathy that is likely to have lower levels of bias. Furthermore, this measure of empathy can be more easily used outside of the research context, hence enhancing the ecological validity of research on empathy. If further tested and used, this new measure could lead to more individuals getting information about their level of empathetic writing. This could allow them to develop their empathy through purposeful interventions on empathy. Family, friends, teachers and counselors might also be able to use this information to interact with individuals more adequately and support their development of empathy. Therefore, this study could play a critical role in starting a research agenda that could promote effective social emotional learning for all.

References

Aggarwal, C. C., & Zhai, C. (2012). A Survey of Text Clustering Algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 77–128). Boston, MA: Springer https://doi.org/10.1007/978-1-4614-3223-4_4

Alexa. (2020). Alexa - Top sites. Retrieved 22 February 2020, from https://www.alexa.com/topsites

American Psychological Association (2020). APA Dictionary of Psychology. Retrieved 16 March 2020, from https://dictionary.apa.org/empathy

Back, M. D., Stopfer. J. M,, Vazire, S., Gaddis, S,, Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science, 21,* 372-374. http://doi.org/10.1177/0956797609360756

Baron-Cohen, S. (1989). The autistic child's theory of mind: A case of specific developmental delay. *Child Psychology & Psychiatry & Allied Disciplines, 30*(2), 285–297. https://doi.org/10.1111/j.1469-7610.1989.tb00241.x

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. & Plumb, I. (2001), The “Reading the Mind in the Eyes” Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, *42*, 241-251. https://doi.org/10.1111/1469-7610.00715

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. *Journal of Autism and Developmental Disorders, 34*(2), 163-175. https://doi.org/10.1023/B:JADD.0000022607.19833.00

Batson, C. D. (1987). Prosocial motivation: Is it ever truly altruistic? In L. Berkowitz (Ed.), *Advances in experimental social psychology, Vol. 20* (pp. 65–122). Cambridge, MA: Academic Press. https://doi.org/10.1016/S0065-2601(08)60412-8

Batt-Rawden, S. A., Chisolm, M. S., Anton, B., & Flickinger, T. E. (2013). Teaching empathy to medical students: An updated, systematic review. *Academic Medicine*, *88*(8), 1171–1177. https://doi.org/10.1097/ACM.0b013e318299f3e3

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin: Springer-Verlag.

Brewer, M. B. (2000). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3–16). Cambridge: Cambridge University Press.

British Educational Research Association (2018). *Ethical Guidelines for Educational*

Research (4th ed., Rep.). London: Author.

Brundrett, M., & Rhodes, C. (2013). *Researching educational leadership and management : methods and approaches*. London: SAGE Publications.

Buechel, S., Buffone, A., Slaff, B., Ungar, L., & Sedoc, J. (2018). Modeling Empathy and Distress in Reaction to News Stories. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4758–4765). Brussels: Association for Computational Linguistics.

Burleson, B. R. (2009). Understanding the outcomes of supportive communication: A dual-process approach. *Journal of Social and Personal Relationships, 26(*1*),* 21–38. https://doi.org/10.1177/0265407509105519

Casale, C., Thomas, C., & Simmons, T. (2018). Developing Empathetic Learners. *Journal of Thought, 52*(3-4), 3-18. https://doi.org/10.2307/90026734

Caskie, G. I. L., Sutton, M. C., & Eckhardt, A. G. (2014). Accuracy of self-reported college GPA: Gender-moderated differences by achievement level and academic self-efficacy. *Journal of College Student Development, 55*(4), 385–390. https://doi.org/10.1353/csd.2014.0038

Császár, I. (2012). *The effect of loving kindness meditation and student teachers stress and empathy* (Doctoral dissertation, Louisiana State University, Baton Rouge, LA). Retrieved from http://etd.lsu.edu/docs/available/etd- 06272012-080110/unrestricted/csaszardiss.pdf

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Davis, M. H. (2015). Empathy and prosocial behavior. In D. A. Schroeder & W. G. Graziano (Eds.), *Oxford library of psychology. The Oxford handbook of prosocial behavior* (pp. 282–306). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399813.013.026

Dimitrijević, A., Hanak, N., Vukosavljević-Gvozden, T., & Opačić, G. (2012). Psychometric properties of the Serbian version of the Empathy Quotient (S-EQ). *Psihologija*, *45*(3), 257–276. https://doi.org/10.2298/PSI1203257D

D'Agostino, R., & Pearson, E. (1973). Tests for Departure from Normality. Empirical Results for the Distributions of b2 and √ b1. *Biometrika, 60*(3), 613-622. https://doi.org/10.2307/2335012

Egan, G. (2014). *The skilled helper: A problem-management and opportunity-development approach to helping*. Belmont, CA: Cengage Learning.

Field, A. P. (2018). *Discovering statistics using IBM SPSS*. London: Sage Publications.

Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York, NY: Basic Books.

Gibson, J., Malandrakis, N., Romero, F., Atkins, D. C., & Narayanan, S. S. (2015). Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. *In Proceedings of INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association* (pp. 1947-1951). Dresden: INTERSPEECH.

Gilbert, E. (2012). Phrases that signal workplace hierarchy. In *Proceedings of 2012 Conference on computer supported cooperative work* (pp. 1037-1046). New York, NY: American Computational Machinery.

Goleman, D. (1998). *Working with emotional intelligence.* New York, NY: Bantam Books.

Greenberg, D. M., Warrier, V., Allison, C., & Baron-Cohen, S. (2018). Testing the Empathizing-Systemizing theory of sex differences and the Extreme Male Brain theory of autism in half a million people. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(48), 12152–12157. https://doi.org/10.1073/pnas.1811032115

Greene, W. H. (2003). *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

Groen, Y., Fuermaier, A. B. M., Den Heijer, A. E., Tucha, O., & Althaus, M. (2015). The Empathy and Systemizing Quotient: The Psychometric Properties of the Dutch Version and a Review of the Cross-Cultural Stability. *Journal of Autism and Developmental Disorders*, *45*(9), 2848–2864. https://doi.org/10.1007/s10803-015-2448-z

Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. Journal of Autism and Developmental Disorders, 24(2), 129–154. https://doi.org/10.1007/BF02172093

Hoffman, M. L. (2000). Empathy and moral development: Implications for caring and justice. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511805851

Hogan, R. (1969). Development of an empathy scale. *Journal of Consulting and Clinical Psychology, 33*, 307–316.

Hojat, M., Vergare, M., Isenberg, G., Cohen, M., & Spandorfer, J. (2015). Underlying construct of empathy, optimism, and burnout in medical students. *International journal of medical education*, *6*, 12–16. https://doi.org/10.5116/ijme.54c3.60cd

Howe, D. (2013). *Empathy: What it is and why it matters.* London: Palgrave Macmillan.

Jensen, K. (2016). Prosociality. *Current Biology*, *26*(16), 748–752. https://doi.org/10.1016/j.cub.2016.07.025

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy

Scale. *Journal of Adolescence*, *29*(4), 589–611. https://doi.org/10.1016/j.adolescence.2005.08.010

Jones, S. M., & Wirtz, J. G. (2006). How Does the Comforting Process Work? An Empirical Test of an Appraisal-Based Model of Comforting. *Human Communication Research*, *32*(3), 217–243. https://doi.org/10.1111/j.1468-2958.2006.00274.x

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing* (2nd ed.). Saddle River, NJ: Prentice-Hall.

Karpathy, A., Johnson, J., & Li, F. (2015). Visualizing and Understanding Recurrent Networks. Retrieved from http://arxiv.org/abs/1506.02078

Kelm, Z., Womer, J., Walter, J. K., & Feudtner, C. (2014). Interventions to cultivate physician empathy: a systematic review. *BMC medical education*, *14*, 219. https://doi.org/10.1186/1472-6920-14-219

Khanpour, H., Caragea, C., & Biyani, P. (2017). Identifying Empathetic Messages in Online Health Communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (pp. 246–251). Taipei: Asian Federation of Natural Language Processing

Kirschenbaum, H. (2013). *Values clarification in counseling and psychotherapy: Practical strategies for individual and group settings.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199972180.001.0001

Kosonogov, V. (2014). The psychometric properties of the Russian version of the empathy quotient. *Psychology in Russia: State of the Art*, *7*(1), 96–104. https://doi.org/10.11621/pir.2014.0110

Lam, T. C. M., Kolomitro, K., & Alamparambil, F. C. (2011). Empathy training: Methods, evaluation practices, and validity. *Journal of Multidisciplinary Evaluation*, *7*, 162–200.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review, 94*(4), 412–426. https://doi.org/10.1037/0033-295X.94.4.412

Litvak, M., Otterbacher, J., Ang, C. S., & Atkins, D. (2016). Social and linguistic behavior and its correlation to trait empathy. In *Workshop on Computational Modeling of People's Opinions, Personality and Emotions in Social Media)*, (pp. 128–137). Osaka: International Conference on Computational Linguistics.

Liu, N. F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 279–290. https://doi.org/10.1080/13562510600680582

Loggia, M. L., Mogil, J. S., & Catherine Bushnell, M. (2008). Empathy hurts: Compassion for another increases both sensory and affective components of pain perception. *Pain*, *136*(1–2), 168–176. https://doi.org/10.1016/j.pain.2007.07.017

Losoya, S. H., & Eisenberg, N. (2001). Affective empathy. In J. A. Hall & F. J. Bernieri (Eds.), *The LEA series in personality and clinical psychology. Interpersonal sensitivity: Theory and measurement* (pp. 21–43). Mahwah, NJ: Lawrence Erlbaum Associates.

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology*, *67*(3), 267–277. https://doi.org/10.1016/j.jclinepi.2013.08.015

Mehl, M. R., & Conner, T. S. (2012). *Handbook of research methods for studying daily life*. New York, NY: Guilford Press.

Melchers, M. C., Li, M., Haas, B. W., Reuter, M., Bischoff, L., & Montag, C. (2016). Similar Personality Patterns Are Associated with Empathy in Four Different Countries. *Frontiers in psychology*, *7*, 290. https://doi.org/10.3389/fpsyg.2016.00290

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from http://arxiv.org/abs/1301.3781

Müller, S., & Moshagen, M. (2019). True virtue, self-presentation, or both?: A behavioral test of impression management and overclaiming. *Psychological Assessment, 31*(2), 181–191. https://doi.org/10.1037/pas0000657

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Neumann, D. L., Chan, R. C. K., Boyle, G. J., Wang, Y., & Westbury, H. R. (2015). Measures of empathy: Self-report, behavioral, and neuroscientific approaches. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (p. 257–289). Cambridge, MA: Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-386915-9.00010-3

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Survey: Computational Sociolinguistics: A Survey. *Computational Linguistics*, *42*(3), 537–593. https://doi.org/10.1162/COLI_a_00258

Palvia, S., Aeron, P., Gupta, P., Mahapatra, D., Parida, R., Rosner, R., & Sindhi, S. (2018). Online Education: Worldwide Status, Challenges, Trends, and Implications. *Journal of Global Information Technology Management*, *21*(4), 233–241. https://doi.org/10.1080/1097198X.2018.1542262

Pushshift. (2020). Reddit Statistics - pushshift.io. Retrieved 22 February 2020, from https://pushshift.io/

Reddit. (2020a). Advice. Retrieved 26 February 2020, from https://www.reddit.com/r/offmychest/

Reddit. (2020b). User Agreement. Retrieved 26 February 2020, from https://www.redditinc.com/policies/user-agreement

Reddit. (2020c). API-terms. Retrieved 26 February 2020, from https://www.reddit.com/wiki/api-terms

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PloS one*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Spreng, R. N., McKinnon, M. C., Mar, R. A., & Levine, B. (2009). The Toronto Empathy Questionnaire: scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of personality assessment*, *91*(1), 62–71. https://doi.org/10.1080/00223890802484381

Srivastava, K., & Das, R. C. (2016). Empathy: Process of adaptation and change, is it trainable?. *Industrial psychiatry journal*, *25*(1), 1–3. https://doi.org/10.4103/0972-6748.196055

Statsmodels, (2020a). statsmodels.discrete.discrete_model.Logit. Retrieved 22 February 2020, from https://www.statsmodels.org/stable/generated/statsmodels.discrete.discrete_model.Logit.html

Statsmodels, (2020b). statsmodels.stats.diagnostic.linear_lm. Retrieved 22 February 2020, from https://www.statsmodels.org/stable/generated/statsmodels.discrete.discrete_model.Logit.html

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Teding van Berkhout, E., & Malouff, J. M. (2016). The efficacy of empathy training: A meta-analysis of randomized controlled trials. *Journal of Counseling Psychology, 63*(1), 32–41. https://doi.org/10.1037/cou0000093

The British Psychological Society. (2014). BPS Code of Human Research Ethics (2nd edition, 2014) | BPS. Retrieved 22 February 2020, from https://www.bps.org.uk/news-and-policy/bps-code-human-research-ethics-2nd-edition-2014

The British Psychological Society. (2017). Ethics Guidelines for Internet-Mediated Research (2017) | BPS. Retrieved 22 February 2020, from https://www.bps.org.uk/news-and-policy/ethics-guidelines-internet-mediated-research-2017

The British Psychological Society. (2018). Code of Ethics and Conduct (2018) | BPS. Retrieved 22 February 2020, from https://www.bps.org.uk/news-and-policy/bps-code-ethics-and-conduct

Upshaw, M. B., Kaiser, C. R., & Sommerville, J. A. (2015). Parents' empathic perspective taking and altruistic behavior predicts infants' arousal to others' emotions. *Frontiers in Psychology, 6,* 360. https://doi.org/10.3389/fpsyg.2015.00360

Verhofstadt, L., Devoldre, I., Buysse, A., Stevens, M., Hinnekens, C., Ickes, W., & Davis, M. (2016). The Role of Cognitive and Affective Empathy in Spouses' Support Interactions: An Observational Study. *PloS one*, *11*(2), e0149944. https://doi.org/10.1371/journal.pone.0149944

Viner, R. M., Russell, S. J., Croker, H., Packer, J., Ward, J., Stansfield, C., … Booy, R. (2020). School closure and management practices during coronavirus outbreaks including COVID-19: a rapid systematic review. *The Lancet Child & Adolescent Health*, *4*(5), 397–404. https://doi.org/10.1016/S2352-4642(20)30095-X

Wakabayashi, A. (2013). Individual differences in empathizing and systemizing in Japanese children: Psychometric properties of the children's versions of the Empathy Quotient (EQ) and Systemizing Quotient (SQ). *Japanese Psychological Research*, *55*(1), 12–19. https://doi.org/10.1111/j.1468-5884.2012.00537.x

Wakabayashi, A., Baron-Cohen, S., Wheelwright, S., Goldenfeld, N., Delaney, J., Fine, D., ... Weil, L. (2006). Development of short forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQ-Short). *Personality and Individual Differences*, *41*(5), 929-940. https://doi.org/10.1016/j.paid.2006.03.017

Warren, C. A. (2018). Empathy, Teacher Dispositions, and Preparation for Culturally Responsive Pedagogy. *Journal of Teacher Education, 69*(2), 169–183. https://doi.org/10.1177/0022487117712487

Weiner, S. J., & Auster, S. (2007). From empathy to caring: defining the ideal approach to a healing relationship. *The Yale journal of biology and medicine*, *80*(3), 123–130.

Wheelwright, S., Baron-Cohen, S., Goldenfeld, N., Delaney, J., Fine, D., Smith, R., Weil, L., & Wakabayashi, A. (2006). Predicting Autism Spectrum Quotient (AQ) from the Systemizing Quotient-Revised (SQ-R) and Empathy Quotient (EQ). *Brain Research, 1079*(1), 47-56. https://doi.org/10.1016/j.brainres.2006.01.012

Wilson, T. D., & Dunn, E. W. (2004). Self-Knowledge: Its Limits, Value, and Potential for Improvement. *Annual Review of Psychology*, *55*(1), 493–518. https://doi.org/10.1146/annurev.psych.55.090902.141954

Xiao B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). "Rate My Therapist": Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. *PLoS ONE*, *10*(12), e0143055. https://doi.org/10.1371/journal.pone.0143055

Yue, X., Jiang, F., Lu, S., & Hiranandani, N. (2016). To Be or Not To Be Humorous? Cross Cultural Perspectives on Humor. *Frontiers in psychology*, *7*, 1495. https://doi.org/10.3389/fpsyg.2016.01495

Zalla, T., & Korman, J. (2018). Prior Knowledge, Episodic Control and Theory of Mind in Autism: Toward an Integrative Account of Social Cognition. *Frontiers in psychology*, *9*, 752. https://doi.org/10.3389/fpsyg.2018.00752

Zhou, Q., Valiente, C., & Eisenberg, N. (2003). Empathy and its measurement. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook of models and measures* (p. 269–284). Worchester, MA: American Psychological Association. https://doi.org/10.1037/10612-017

Zickfeld, J. H., Schubert, T. W., Seibt, B., & Fiske, A. P. (2017). Empathic Concern Is Part of a More General Communal Emotion. *Frontiers in psychology*, *8*, 723. https://doi.org/10.3389/fpsyg.2017.00723

Appendix A: Empathy Quotient Questionnaire

Empathy Questionnaire

This questionnaire comprises of your Reddit username and 22 items, which take around 2 mins.

Give the first answer that comes to your mind for accuracy! Your gender and age are optional.

* Required

Details

The questionnaire is based on Wakabayashi et al., 2006 [available here at http://guava.physics.uiuc.edu/~nigel/REPRINTS/2006/Wakabayashi%20Development%20of%20short%20forms%20of%20the%20Empathy%20PerIndDiff%202006%20(PDF).pdf]

Your answers will not be recorded until the submit button is clicked and you are allowed to withdraw from the study any time during the survey. You may also choose to withdraw from this process after submitting your survey by sending a message to the Reddit account (confused_doo_doo). By clicking submit, you agree to allow us to use the submitted data for research purposes.

As part of my research, data collected will only be used to study the correlation between word use and Empathy scores at the population level rather than for any individual. Personal data collected will be stored securely based on GDPR and access will only be provided to the research team.

Reddit Username *

Your answer

Empathy Questionnaire * 33 points

	strongly agree	agree	disagree	strongly disagree
I can easily tell if someone else wants to enter a conversation.	○	○	○	○
I really enjoy caring for other people.	○	○	○	○
I find it hard to know what to do in a social situation	○	○	○	○
I often find it difficult to judge if something is rude or polite	○	○	○	○
In a conversation, I tend to focus on my own thoughts rather than on what my listener might be thinking	○	○	○	○
I can pick up quickly if someone says one thing but means another.	○	○	○	○
It is hard for me to see why some things upset people so much.	○	○	○	○
I find it easy to put myself in somebody else's shoes.	○	○	○	○
I am good at predicting how someone will feel.	○	○	○	○

I am quick to spot when someone in a group is feeling awkward or uncomfortable.	○	○	○	○
I can't always see why someone should have felt offended by a remark.	○	○	○	○
I don't tend to find social situations confusing.	○	○	○	○
Other people tell me I am good at understanding how they are feeling and what they are thinking.	○	○	○	○
I can easily tell if someone else is interested or bored with what I am saying.	○	○	○	○
Friends usually talk to me about their problems as they say that I am very understanding.	○	○	○	○
I can sense if I am intruding, even if the other person doesn't tell me.	○	○	○	○
Other people often say that I am insensitive, though I don't always see why	○	○	○	○
I can tune into how someone	○	○	○	○

else feels rapidly and intuitively.				
I can easily work out what another person might want to talk about.	○	○	○	○
I can tell if someone is masking their true emotion.	○	○	○	○
I am good at predicting what someone will do.	○	○	○	○
I tend to get emotionally involved with a friend's problems.	○	○	○	○

Gender

- ○ Male
- ○ Female
- ○ Prefer not to say/Others

Age

Your answer

Page 1 of 1 Submit

Appendix B: Details of ensuring comments are independent from one another

Comments on the same post might not be independent since they might address similar issues mentioned by the post. Furthermore, the probability that a comment will be labeled as helpful depends on the post, as shown in Figure B.1. To overcome this, only one comment was randomly chosen from each post to be included.

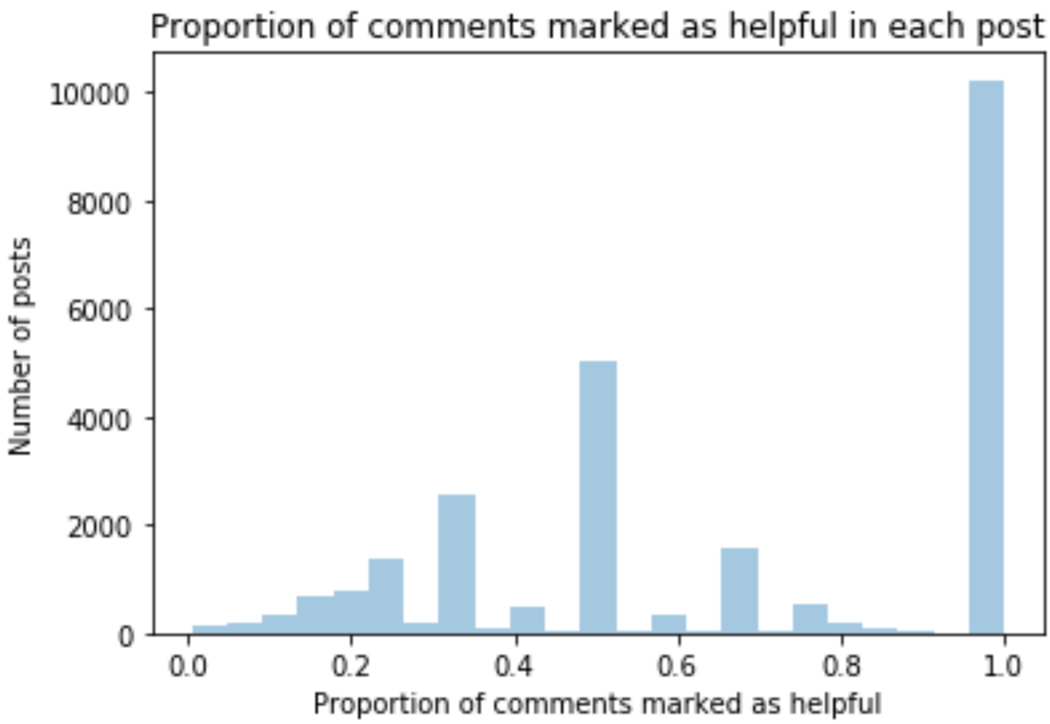

Figure B.1 Proportion of comments marked as helpful in each post.

Comments from the same user also might not be independent because the same user might have similar patterns of word use across comments. This is evident in differences in the proportion of comments marked as helpful from each user in Figure B.2. This suggests stable patterns of word use of users leads to different rates of being helpful. To overcome this, only one comment from each user was included.

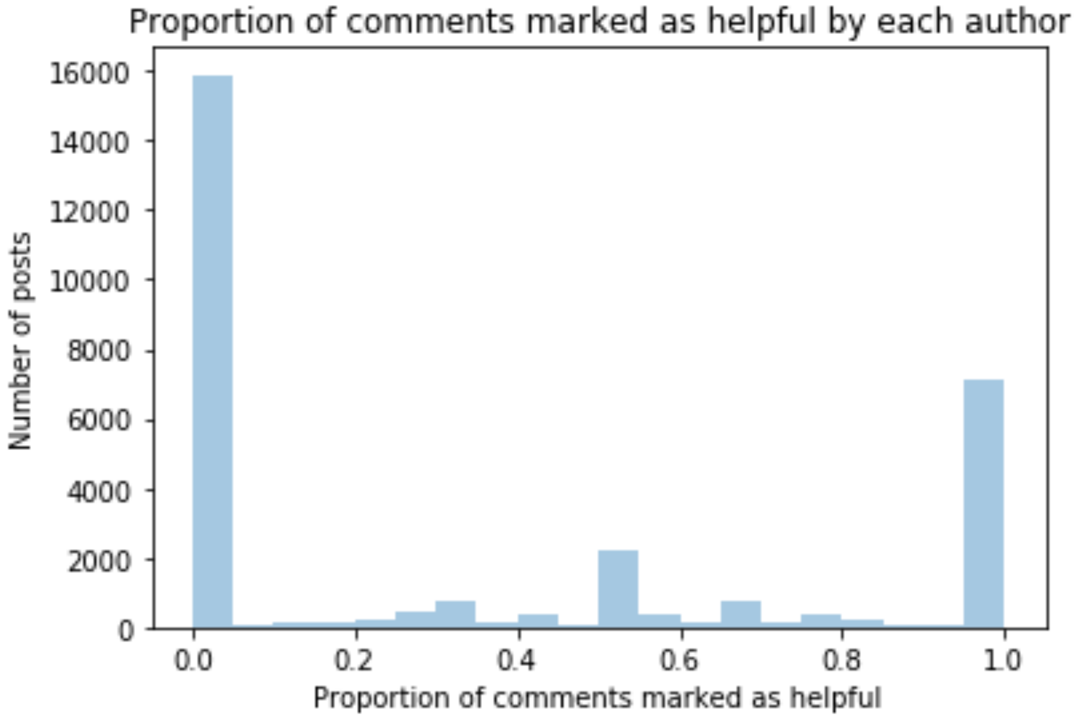

Figure B.2. Proportion of comments marked as helpful from each user.

Appendix C: Checking assumptions of logistic and linear regressions

Linear regression between the proportion of helpful comments and Empathy Quotient score

Linearity

To check for linearity, the Lagrange Multiplier test (Statsmodels, 2020b) for the correctness of the linear specification was conducted. The hypothesis that the linear specification is correct cannot be rejected ($\lambda = 2.01, two - tailed\ p < 0.1612$).

Normal distribution of residuals

Based on a two-tailed chi-squared test on the sum of normalized kurtosis squared and normalized skew squared (Agostino & Pearson, 1973),
the hypothesis that residuals are normally distributed cannot be rejected ($\chi^2(2, N = 91) = 1.915\ , two - tailed\ p = 0.3838$).

Homoscedasticity

Based on the Breusch-Pagan Lagrange Multiplier test for heteroscedasticity (Greene, 2003), the hypothesis that the residual does not vary based on the independent variable can be rejected ($\lambda = 36.48, two - tailed\ p < 0.001$).

Logistic and linear regression for word use against effective helpfulness of comments and users respectively

These two regressions are discussed together since they shared many similarities.

Multi-collinearity

First, independent variables need to have low multi-collinearity. Because of the large number of independent variables, it was not computationally feasible to calculate the Variance Inflation Factor (VIF) (Field, 2018) for each independent variable, alongside related research (Schwartz et al, 2013). However, the VIF was calculated for each significant predictor. Nearly all VIFs (M=2.678, SD= 3.281) are below 10, suggesting low multi-collinearity (Field, 2018).

Large sample size

While a large sample size is recommended in typical logistic regressions, logistic regressions involving text data generally have a large number of independent variables relative to the number of samples (Jurafsky & Martin, 2009). This is acceptable in text-related logistic regression (Aggarwal & Zhai, 2012; Schwartz et al, 2013; Gilbert, 2012) because while individual words might only contribute a small effect to the prediction, the summative ability of all words to influence prediction is large due to the high number of words in each comment. This is also observed in the findings of this study.

Linearity (Linear Regression only)

To check for linearity, the Lagrange Multiplier test (Statsmodels, 2020b) for the correctness of the linear specification was conducted. The hypothesis that the linear specification is correct can be rejected ($\lambda = 16143, two - tailed\ p < 0.001$). However, limitations in computational resources means that non-linear transformations of variables can only be included in future work.

Normal distribution of residuals (Linear Regression only)

Based on a two-tailed chi-squared test on the sum of normalized kurtosis squared and normalized skew squared (Agostino & Pearson, 1973), the hypothesis that residuals are normally distributed can be rejected ($\chi^2(2, N = 50800) = 4478, two - sided\ p < 0.001$). This is expected given that a linear regression is not the optimal model and the more appropriate fractional logit regression model (Statsmodels, 2020a) will be used in the future given more adequate computational resources.

Homoscedasticity (Linear Regression only)

Heteroscedasticity concerning single independent variables is acceptable for language related regressions (Jurafsky & Martin, 2009) because for each independent variable/word, most entries are zero (most text will not contain a particular word, except when the word is commonly used). Therefore, it is expected that the residual at word-count = 0 to have much greater variance than other levels of word count.

Appendix D: Ethical challenges and solutions

A possible challenge was whether consent was necessary in utilizing comments and posts written by Reddit users. Based on BERA (2018), online discussion forums constitute data that were explicitly produced for public use, meaning that consent was not required. Similarly, BPS (2017) proposed similar guidance to Internet data in the 'public domain' – previously defined as public behavior that 'would be expected to be observed by strangers' (BPS, 2014, p.25). Since Reddit posts and comments are readily accessible to anyone on the Internet, they undoubtedly are in the 'public domain'. Guided by BPS (2017), the End User License Agreement (Reddit, 2020b) and Terms of Use for Developers (Reddit, 2020c) were also looked into. They show that Reddit Users grant permission for their written contents to be used by Reddit and its developer partners upon signing up for the service. Finally, the public use of Reddit is also uncurbed by the General Data Protection Regulation (GDPR) since Reddit does not contain personally identifiable information.

Another potential challenge arose from the collection of information through an online questionnaire (available in appendix A). To prevent dealing with sensitive information from minors, recruitment of participants to answer the questionnaire was restricted to those aged 18 and above. The questionnaire used was adopted from existing research (Wakabayashi et al., 2006), and the items do not concern any sensitive areas. Expectedly, no harm was reported by participants in Wakabayashi et al. (2006) or this study. In accordance to BERA (2018), participants were guided to make an informed decision about their participation, giving them the opportunity to withdraw at any time, during and after the completion of the questionnaire. Finally, collected information, which does not include personally identifiable information, was stored securely in accordance to GDPR with access restricted to the research investigator.

Appendix E: Significant predictors of effective helpfulness

Word	Beta	Standard Error	t-statistic	One-tailed p
glad	1.0588	0.3051	3.4708	0.0003
welcome	1.1116	0.366	3.0371	0.0012
try	0.2532	0.0883	2.8671	0.0021
you	0.0746	0.0261	2.8589	0.0021
form	1.5547	0.5573	2.7898	0.0026
luck	0.4135	0.1611	2.5663	0.0051
date	0.6013	0.2354	2.5543	0.0053
hopefully	0.954	0.3773	2.5283	0.0057
let	0.3129	0.1244	2.5151	0.0059
doesnt	1.2358	0.5161	2.3943	0.0083
for	0.1185	0.0501	2.3629	0.0091
friend	0.2349	0.105	2.236	0.0127
yes	0.4628	0.2125	2.1782	0.0147
forward	0.8802	0.4103	2.1451	0.0160
other	0.2582	0.1223	2.1114	0.0174
takes	0.7625	0.3627	2.1023	0.0178
lives	0.8772	0.4182	2.0975	0.0180
sometimes	0.3966	0.1895	2.0933	0.0182
anymore	0.654	0.3133	2.0876	0.0184
feels	0.5594	0.2691	2.079	0.0188
strong	0.7325	0.3546	2.0658	0.0194
it	0.0643	0.0315	2.0369	0.0208
give	0.2648	0.1314	2.0159	0.0219
np	1.2397	0.6191	2.0024	0.0226
ago	0.7153	0.3573	2.0019	0.0226
down	0.3438	0.172	1.9988	0.0228
hope	0.3179	0.1614	1.97	0.0244
learned	0.9197	0.4677	1.9665	0.0246
good	0.1706	0.088	1.9388	0.0263
personally	0.4906	0.2536	1.9346	0.0265
just	0.1036	0.0538	1.9249	0.0271
stories	0.9211	0.4897	1.8811	0.0300
get	0.1301	0.0692	1.88	0.0301
coming	0.6291	0.3357	1.8738	0.0305
happens	0.5182	0.2767	1.873	0.0305
24	1.2901	0.7001	1.8426	0.0327
would	0.133	0.0727	1.8301	0.0336
probably	0.2511	0.1392	1.8042	0.0356
brain	0.7009	0.3912	1.7917	0.0366
value	0.8304	0.4666	1.7796	0.0376
likely	0.4768	0.2709	1.7596	0.0392
hey	0.4594	0.2632	1.7454	0.0405
helped	0.3595	0.2076	1.7319	0.0416

till	0.8273	0.4796	1.725	0.0423
again	0.318	0.1861	1.7094	0.0437
careful	0.7423	0.4396	1.6888	0.0456
taught	1.2071	0.7227	1.6703	0.0474
swear	1.2041	0.721	1.67	0.0475
cover	1.1195	0.6704	1.67	0.0475
comfortable	0.4385	0.2632	1.6663	0.0478

Table E.1 Significant positive predictors of effective helpfulness of comments

Word	Beta	Standard Error	t-statistic	One-tailed p
told	-0.6697	0.2422	-2.7656	0.0028
taken	-1.1331	0.4449	-2.5467	0.0054
victim	-2.1648	0.8654	-2.5016	0.0062
read	-0.5985	0.2426	-2.4673	0.0068
dating	-0.6031	0.26	-2.3193	0.0102
whenever	-1.1293	0.4918	-2.2965	0.0108
oh	-0.6653	0.3111	-2.1386	0.0162
world	-0.5489	0.2587	-2.1221	0.0169
trans	-0.8937	0.4355	-2.0523	0.0201
advice	-0.3141	0.1534	-2.0481	0.0203
dealt	-1.7782	0.8689	-2.0465	0.0204
until	-0.4142	0.205	-2.0202	0.0217
fucking	-0.7229	0.3642	-1.9849	0.0236
tips	-1.535	0.7833	-1.9596	0.0250
interests	-0.687	0.3539	-1.9412	0.0261
must	-0.7678	0.4031	-1.905	0.0284
through	-0.2796	0.1477	-1.8932	0.0292
wearing	-1.3322	0.7097	-1.8773	0.0302
asking	-0.4822	0.2607	-1.85	0.0322
compare	-1.5334	0.8297	-1.8482	0.0323
interested	-0.4952	0.2689	-1.8418	0.0328
gender	-1.039	0.5645	-1.8404	0.0329
leading	-1.7162	0.943	-1.82	0.0344
lead	-0.8828	0.489	-1.8053	0.0355
his	-0.1984	0.1101	-1.8018	0.0358
keep	-0.2393	0.1338	-1.7884	0.0369
engage	-1.462	0.8277	-1.7664	0.0387
happen	-0.3797	0.2152	-1.7643	0.0388
level	-0.8604	0.4901	-1.7554	0.0396
kill	-0.8811	0.5075	-1.7363	0.0413
easy	-0.465	0.2687	-1.7303	0.0418
bad	-0.2541	0.1475	-1.7229	0.0425
history	-0.8716	0.5063	-1.7217	0.0426
cutting	-1.0409	0.6065	-1.7163	0.0431
communicati	-0.8371	0.4878	-1.716	0.0431

on				
put	-0.2931	0.1713	-1.7116	0.0435
depression	-0.4266	0.2495	-1.7098	0.0437
woman	-0.6516	0.3861	-1.6877	0.0457
rid	-0.9584	0.5697	-1.6823	0.0463
random	-0.761	0.4526	-1.6813	0.0464
see	-0.1875	0.1118	-1.6763	0.0468
promise	-0.75	0.4491	-1.6701	0.0475
op	-0.4125	0.249	-1.6565	0.0488
wish	-0.4175	0.2535	-1.6472	0.0498

Table E.2 Significant negative predictors of effective helpfulness of comments

Word	Beta	Standard Error	t-statistic	One-tailed p
nta	0.2951	0.0124	23.7147	0.0000
thank	0.2623	0.0135	19.3726	0.0000
thanks	0.2518	0.0134	18.7616	0.0000
yeah	0.1852	0.0104	17.8019	0.0000
it	0.0321	0.0019	17.0428	0.0000
yta	0.2949	0.0181	16.3151	0.0000
welcome	0.2714	0.02	13.5922	0.0000
yup	0.4375	0.0332	13.1583	0.0000
yes	0.1326	0.0102	13.0382	0.0000
removed	0.2436	0.0188	12.9384	0.0000
yep	0.3106	0.0268	11.6009	0.0000
no	0.0554	0.005	11.1116	0.0000
nah	0.2269	0.0209	10.857	0.0000
glad	0.2029	0.0189	10.7176	0.0000
sorry	0.1151	0.0111	10.3511	0.0000
agree	0.1423	0.0138	10.324	0.0000
lol	0.1125	0.011	10.1994	0.0000
sounds	0.0909	0.0098	9.2998	0.0000
go	0.0511	0.0055	9.2877	0.0000
hahaha	0.3494	0.0384	9.0865	0.0000
what	0.0362	0.004	9.0044	0.0000
good	0.0516	0.0058	8.8916	0.0000
beautiful	0.2129	0.0243	8.7588	0.0000
re	0.0367	0.0044	8.2999	0.0000
try	0.0588	0.0072	8.1534	0.0000
depends	0.1555	0.0191	8.1502	0.0000
looks	0.1158	0.0145	8.0007	0.0000
why	0.0537	0.0068	7.9397	0.0000
love	0.0674	0.0087	7.792	0.0000
have	0.0244	0.0032	7.5525	0.0000
definitely	0.0948	0.0126	7.5377	0.0000
impressive	0.3389	0.0462	7.3358	0.0000

think	0.0366	0.005	7.2915	0.0000
get	0.0305	0.0043	7.1423	0.0000
troll	0.2279	0.0323	7.0539	0.0000
just	0.0248	0.0035	7.0238	0.0000
do	0.0263	0.0038	6.9794	0.0000
ouch	0.3962	0.0569	6.9637	0.0000
talk	0.0601	0.0087	6.9382	0.0000
maybe	0.0493	0.0071	6.9028	0.0000
youtu	0.3003	0.0439	6.8387	0.0000
esh	0.2711	0.0401	6.7608	0.0000
haha	0.1652	0.0244	6.7551	0.0000
bloop	1.104	0.1643	6.7201	0.0000
meme	0.2488	0.0375	6.6444	0.0000
report	0.1228	0.0185	6.6369	0.0000
lmao	0.1938	0.0292	6.6311	0.0000
expelled	0.5909	0.0902	6.5513	0.0000
personalfinance	0.4282	0.0655	6.5364	0.0000
move	0.071	0.0109	6.5249	0.0000

Table E.3 50 Most significant positive predictors of effective helpfulness of users

Word	Beta	Standard Error	t-statistic	One-tailed p
and	-0.0458	0.002	-23.3578	0.0000
to	-0.0281	0.002	-14.2213	0.0000
but	-0.0415	0.0035	-11.7658	0.0000
because	-0.0563	0.0055	-10.1976	0.0000
gt	-0.0781	0.0082	-9.5654	0.0000
we	-0.0409	0.0048	-8.5537	0.0000
or	-0.0276	0.0035	-7.8396	0.0000
even	-0.0451	0.0066	-6.8744	0.0000
etc	-0.0919	0.0135	-6.8096	0.0000
which	-0.06	0.0089	-6.7198	0.0000
bleep	-1.3963	0.2121	-6.5841	0.0000
if	-0.0205	0.0032	-6.4456	0.0000
footer	-0.6636	0.1079	-6.1474	0.0000
lastly	-0.5214	0.0986	-5.2862	0.0000
example	-0.113	0.0221	-5.1146	0.0000
tl	-0.4712	0.0974	-4.8386	0.0000
when	-0.0257	0.0053	-4.7999	0.0000
obligations	-0.6627	0.1392	-4.7609	0.0000
of	-0.0131	0.0028	-4.6507	0.0000
while	-0.0456	0.01	-4.574	0.0000
piano	-0.4805	0.107	-4.4899	0.0000
yr	-0.3202	0.0718	-4.4593	0.0000
life	-0.0302	0.0069	-4.378	0.0000

unaware	-0.4495	0.1027	-4.3769	0.0000
wears	-0.5288	0.1243	-4.253	0.0000
as	-0.0155	0.0037	-4.1696	0.0000
quad	-0.5458	0.1314	-4.1542	0.0000
reasoning	-0.3132	0.0756	-4.1406	0.0000
wit	-0.4494	0.1087	-4.1356	0.0000
petition	-0.25	0.0605	-4.1305	0.0000
statue	-0.2632	0.0638	-4.1283	0.0000
gathered	-0.6258	0.1536	-4.075	0.0000
lengths	-0.493	0.1226	-4.0201	0.0000
tact	-0.6297	0.1573	-4.0033	0.0000
scam	-0.3952	0.099	-3.9905	0.0000
restrictions	-0.3735	0.0945	-3.9525	0.0000
will	-0.0176	0.0045	-3.9331	0.0000
beyond	-0.1361	0.0349	-3.9057	0.0000
pet	-0.1007	0.0261	-3.8502	0.0001
vitamins	-0.3217	0.0842	-3.8223	0.0001
perspective	-0.1193	0.0317	-3.7672	0.0001
comparing	-0.232	0.0618	-3.7515	0.0001
sis	-0.3277	0.0879	-3.7296	0.0001
managing	-0.2835	0.076	-3.7295	0.0001
intake	-0.2811	0.0756	-3.7178	0.0001
leading	-0.2262	0.0612	-3.6934	0.0001
masters	-0.2538	0.0692	-3.6687	0.0001
clue	-0.2403	0.0661	-3.6374	0.0001
pulling	-0.2039	0.0562	-3.6275	0.0001
someway	-0.4343	0.1203	-3.6097	0.0002

Table E.4 50 Most significant negative predictors of effective helpfulness of users